



Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Math

2015-2016 School Year

February 24, 2017

ARIZONA STATEWIDE ASSESSMENTS

**ARIZONA'S MEASUREMENT OF EDUCATIONAL READINESS TO INFORM
TEACHING (AzMERIT)**

ELA GRADES 3-11

MATH GRADES 3-8, ALGEBRA I, GEOMETRY, AND ALGEBRA II

2015-2016 ANNUAL TECHNICAL REPORT

FEBRUARY 24, 2017

Prepared by American Institutes for Research (AIR) in collaboration with the
Arizona Department of Education

TABLE OF CONTENTS

1.	Introduction: The Validity of AzMERIT Test Score Interpretations.....	9
1.1	Overview	9
1.2	Validity Evidence	10
1.3	Evidence Based on Test Content	18
1.4	Evidence for Interpretation of Performance Standards.....	20
1.5	Evidence Based on Internal Structure	23
1.5.1	ELA Content Model	25
1.5.2	ELA Depth of Knowledge.....	26
1.5.3	Math Content Model.....	27
1.5.4	Math Depth of Knowledge	28
1.6	Evidence for Relationships with Conceptually Related Constructs.....	29
1.7	Measurement Invariance Across Subgroups.....	30
1.8	Differential Mode Effects Across subgroups	32
1.9	Evidence for Student Growth – Overall and by subgroups	33
1.10	Day, week, and time of day effects on performance	36
1.11	Summary of Validity of Test Score Interpretations.....	37
2.	Background of Arizona Statewide Assessments	39
2.1	Development of Arizona College and Career Ready Standards (ACCRS).....	40
2.2	AzMERIT Test Design	40
3.	Summary of Fall 2015 Operational Test Administration	42
3.1	Student Population and Participation	42
3.2	Summary of Overall Student Performance	43
3.3	Student Performance by Subgroup.....	44
3.4	Reliability.....	46
3.4.1	Internal Consistency	47
3.4.2	Standard Error of Measurement	47

3.4.3	Student Classification Reliability	50
3.4.4	Classification accuracy	50
3.4.5	Classification consistency.....	51
3.4.6	Classification reliability estimates.....	51
3.4.7	Reliability for Sub-Groups in the Population.....	52
3.4.8	Subscale Reliability.....	53
3.5	Subscale Intercorrelations.....	53
4.	Summary of Spring 2016 Operational Test Administration.....	55
4.1	Student Population and Participation	55
4.2	Classical Item Analysis	56
4.3	Item Response Theory Analysis.....	58
4.4	Summary of Overall Student Performance	61
4.5	Student Performance by Subgroup.....	63
4.6	Reliability.....	68
4.6.1	Internal Consistency.....	69
4.6.2	Standard Error of Measurement	69
4.6.3	Student Classification Reliability	73
4.6.4	Classification Accuracy	74
4.6.5	Classification Consistency	74
4.6.6	Classification Accuracy and Consistency Estimates	75
4.6.7	Reliability for Sub-Groups in the Population.....	81
4.6.8	Subscale Reliability.....	83
4.7	Subscale Intercorrelations.....	85
4.8	Rater Effects	87
5.	Item Development & Test Construction	88
5.1	Item Development Process	89
5.1.1	Item Writing	89

5.1.2	Machine-Scored Constructed-Response Item Development Tools	92
5.1.3	Item Types.....	93
5.2	Item Review.....	94
5.3	Field Testing	96
5.4	Item Statistics.....	97
5.4.1	Classical Statistics.....	97
5.4.2	IRT Stats.....	98
5.4.3	Analysis of Differential Item Functioning	98
5.5	Test Construction	100
5.5.1	Operational Form Construction	100
5.5.2	Assembling Test Forms	101
6.	Test Administration	103
6.1	Eligibility	103
6.2	Administration Procedures	103
6.3	Testing Conditions, Tools, and Accommodations	106
6.3.1	Universal Test Administration Conditions.....	106
6.3.2	Universal Testing Tools for Computer Based Testers.....	107
6.3.3	Subject Area Tools for CBT and PBT	108
6.3.4	Accommodations	109
6.4	System Security	113
6.4.1	Secure System Design	113
6.4.2	System Security Components.....	113
6.5	Test Security	115
6.6	Data Forensics Program	117
6.6.1	Changes in Student Performance.....	117
6.6.2	Item Response Latency	118
6.6.3	Inconsistent Item Response Pattern (Person Fit).....	118

6.6.4	Response Change and Response Similarity.....	119
7.	Reporting and Interpreting AzMERIT Scores	121
7.1	Appropriate Uses for Scores and Reports	121
7.2	Reports Provided.....	122
7.2.1	Family Reports.....	122
7.2.2	Online Reporting System for Educators	123
7.3	Interpretation of Scores	127
8.	Performance Standards	129
8.1	Standard Setting Procedures.....	129
8.1.1	Performance Level Descriptors	130
8.2	Recommended Performance Standards	130
9.	Scaling And Equating	134
9.1	Item Response Theory Procedures	135
9.1.1	Calibration of AzMERIT Item Banks.....	135
9.1.2	Estimating Student Ability Using Maximum Likelihood Estimation	136
9.2	Establishing a Vertical Scale in ELA and Math	137
9.2.1	Linking Items	137
9.2.2	Linking Analysis	138
9.3	AzMerit Reporting Scale (Scale Scores).....	142
9.4	Linking paper and Online Test Scores (Mode Comparability)	143
9.4.1	Mode Linking.....	143
9.4.2	School Performance	147
9.5	Linking the AzMERIT to Other Scales for Performance Comparison.....	147
9.5.1	Establishing Linkages to AIMS, SAGE, Smarter Balanced, PISA	147
9.5.2	Identifying the Location of the ACT College-Ready Cut on AzMERIT	148
10.	Constructed-Response Scoring.....	151
10.1	Machine Scoring.....	151

10.1.1	Explicit Rubrics	151
10.1.2	Essay Autoscoring	151
10.2	Hand-Scoring	164
10.2.1	Handscoring Process	164
10.2.2	Hand-Scoring Quality Control	165
10.2.3	Hand-Scoring Reliability and Validity	165
10.2.4	Machine-Scoring verification	167
11.	Quality Assurance Procedures.....	168
11.1	Quality Assurance in Test Construction	168
11.2	Quality Assurance in Paper-Delivered Test Production	169
11.3	Quality Assurance in Computer-Delivered Test Production	170
11.3.1	Production of Content	170
11.3.2	Web Approval of Content During Development.....	170
11.3.3	Approval of Final Forms	171
11.3.4	Packaging	171
11.3.5	Platform Review.....	171
11.3.6	User Acceptance Testing and Final Review.....	171
11.3.7	Functionality and Configuration	172
11.4	Quality Assurance in Document Processing.....	172
11.4.1	Scanning Accuracy	172
11.4.2	Quality Assurance in Editing and Data Input	172
11.5	Quality Assurance in Data Preparation	173
11.6	Quality Assurance in Test Form Equating	174
11.7	Quality Assurance in Scoring and Reporting	174
11.7.1	Quality Assurance in Hand Scoring	174
11.7.2	Test Scoring.....	176
11.7.3	Reporting	178

12.	References.....	180
-----	-----------------	-----

APPENDICES

Appendix A. AzMERIT ELA and Mathematics Test Blueprints	A-1
Appendix B. Measurement Invariance Testing by Subgroups	B-1
Appendix C. Regression Model Parameter Estimates of Differential Growth across Subgroups.....	C-1
Appendix D. Student Participation by Demographic Subgroup – Fall 2015 Administration	D-1
Appendix E. Equations and Formula for Estimating Reliability.....	E-1
Appendix F. Standard Errors of Measurement – Fall 2015 Administration.....	F-1
Appendix G. Student Participation by Demographic Subgroup – Spring 2016 Administration.....	G-1
Appendix H. Operational Item Parameter Estimates – Spring 2016 Administration.....	H-1
Appendix I. Standard Errors of Measurement – Spring 2016 Administration	I-1
Appendix J. ELA Writing Prompt Rater Agreement Report – Spring 2016 Administration	J-1
Appendix K. Linking Spring 16 Operational Parameters to the Bank Scale	K-1
Appendix L. Data Review Training Slides	L-1
Appendix M. Test Characteristic Curves – Spring 2016 Administration	M-1

1. INTRODUCTION: THE VALIDITY OF AZMERIT TEST SCORE INTERPRETATIONS

1.1 OVERVIEW

The purpose of this technical report is to document the evidence supporting the claims made for how AzMERIT test scores may be interpreted. Evidence for the validity of test score interpretations is central to claims that AzMERIT test scores can be used to evaluate the effectiveness with which Arizona districts and schools teach students the Arizona College and Career Ready Standards and whether individual students have achieved those standards by the end of each school year. Thus, the report begins with a review of validity evidence evaluated to date. Evidence for the validity of test score interpretations is expected to accrue over time, so that this section will be expanded as further evidence is gained.

Chapter 2 of the report describes the design and development of the AzMERIT assessment system, including the Arizona College and Career Ready standards which define the content domain to be assessed by AzMERIT, the development of test specifications, including blueprints, that ensure the breadth and depth of the content domain is adequately sampled by the assessments, as well as test development procedures that ensure alignment of test forms with the blueprint specifications.

Chapters 3 and 4 provide summaries of the AzMERIT test administrations. Chapter 3 presents results of the fall 2015 administration of the high school end-of-course (EOC) assessments, and Chapter 4 presents results of the spring 2016 administration of the full AzMERIT assessment system, including end of year assessments in ELA and math for grades 3-8, and high school EOC tests in ELA and math. These chapters provide summaries of the test taking student population and their performance on the assessments. In addition, these chapters describe administration specific evidence for the reliability of the AzMERIT assessments, including internal consistency reliability, standard errors of measurement, and the reliability of performance level classifications.

The remaining chapters document technical details of the test development, administration, scoring, and reporting activities. Chapter 5 describes the item development process and especially the sequence of reviews that each item must pass through before being eligible for AzMERIT test administration. This chapter also describes the procedures for constructing test forms from items successfully passing through the review process.

Chapter 6 documents the test administration procedures, including eligibility of participation in the AzMERIT assessments, testing conditions, including accessibility tools and accommodations, systems security for assessments administered online, as well as test security procedures for all test administrations.

A description of the score reporting system and the interpretation of test scores is provided in Chapter 7. Chapter 8 describes the procedures that the Arizona Department of Education (ADE) used to identify and adopt performance standards for AzMERIT assessments, and Chapter 9 describes the procedures used to scale and equate the AzMERIT assessments for scoring and reporting.

Chapter 10 describes the procedures for scoring constructed-response items, both machine- and hand-scored items, and provides summary rater agreement results. Finally, Chapter 11 provides an overview of the quality assurance processes described throughout that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.2 VALIDITY EVIDENCE

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support the validity of test score interpretations.

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the standards make explicit that validity is not an attribute of tests, but rather of test score interpretations. Some test score interpretations may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Central to evaluating the validity of test score interpretations is determining whether the test measures the intended construct. Such an evaluation in turn requires a clear definition of the measurement construct. For Arizona's AzMERIT, the definition of the measurement construct is provided by the [Arizona College and Career Ready Standards \(ACCRS\)](#).

In 2010, Arizona adopted new academic content standards in English language arts (ELA) and math. The Arizona College and Career Ready Standards are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate.¹ In spring 2015, the ADE administered Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) to assess proficiency on the new Arizona College and Career Ready Standards for the first time. The AzMERIT measures English language arts (ELA) and math in grades 3-8 and following completion of high school coursework in ELA Grade 9, ELA Grade 10, ELA Grade 11, Algebra I, Geometry, and Algebra II.

Because directly measuring student achievement against each benchmark in the Arizona College and Career Ready Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the ACCRS.² To ensure that each student is assessed on the intended breadth and depth of the ACCRS, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark.³ Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand

¹ Standard 1.1 – The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

² Standard 4.0 – Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

³ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzMERIT [ELA](#) and [math](#) test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels. The ELA and math blueprints are also provided as an attachment in Appendix A.

While the blueprints ensure that the full range of the intended measurement construct is represented in each test administration, tests may also inadvertently measure attributes that are not relevant to the construct of interest. For example, when a high level of English language proficiency is necessary to access content in other subject area assessments such as math or science, language proficiency may unnecessarily limit the student's ability to demonstrate achievement in those subject areas. Thus, while such tests may measure achievement of relevant subject area content standards, they may also measure construct irrelevant variation in language proficiency, limiting the generalizability of test score interpretations for some student populations.

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement.⁴ Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including items and accompanying stimuli. In the review process, adherence to the principles of universal design is verified.

In addition, the AzMERIT test delivery system provides a range of accessibility tools and accommodations for reducing construct irrelevant barriers to accessing test content for virtually all students.⁵ The range of accommodations, provided in the online testing environment, far exceed the typical accommodations made available in paper-based test administrations. Exhibits 1.2.1-1.2.5 list the accommodations and accessibility supports currently available for students taking the AzMERIT assessments online. Paper test forms are available as

⁴ Standard 3.0 – All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

⁵ Standard 3.1 – Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.

Standard 3.2 – Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

Standard 12.3 – Those responsible for the development and use of educational assessments should design all relevant steps of the testing process to promote access to the construct for all individuals and subgroups for whom the assessment is intended.

an accommodation for students testing in online schools should the accommodations provided online not be sufficient to remove barriers to accessing test content. These include both larger print and Braille forms, which are also available, for students who need them, in schools administering AzMERIT as a paper-based assessment. Section 6.3 describes available testing tools and accommodations for students testing online and on paper.

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment. Universal test administration conditions are available for both paper-based test (PBT) and computer-based testing (CBT) modes. Universal test administration conditions include:

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- Extended time. (Testing session must be completed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan as an accommodation, for AzMERIT testing purposes, these are not considered testing accommodations and are available to any student who needs them not just to students with IEPs.

Exhibit 1.2.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 1.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to click anywhere on the selected response text or button for multiple choice options.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen <i>Test Instructions and Help</i> .
Highlighter	Highlight text in a passage or item.
Line Reader	Allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	Allows the session to be paused at any time and restarted and taken over a one day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 1.2.2.

Exhibit 1.2.2 Subject Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	<p>CBT – Students have access to the dictionary/thesaurus tool. Students may opt to use a published, paper dictionary or thesaurus instead of using this tool.</p> <p>PBT – Schools must make published, paper dictionaries and thesauruses available to students.</p> <p>Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned-off.</p>
Writing Guide	Writing	<p>CBT – Students have access to the writing guide tool.</p> <p>PBT – The writing guide is included within the test booklet.</p>
Scratch Paper	Writing and Math	<p>CBT – Schools must provide scratch paper (plain, lined, or graph) to students.</p> <p>PBT – Schools must provide scratch paper (plain, lined, or graph) to students.</p>
<p>Calculator</p> <p>Grades 7-8 (Part 1 only): scientific calculators are acceptable</p> <p>EOC (entire test): graphing calculators are acceptable</p>	Math	<p>CBT – Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted.</p> <p>PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.</p>

Accommodations are provisions made in how a student accesses and demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Language Learner or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education need, or language need and the accommodation(s)

provided to the student during educational activities, including assessment. Test administrators are instructed to make accommodation decisions based on individual needs, and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzMERIT test that is not already used regularly in the classroom.

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section. Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student’s individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact ADE for guidance.

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. There are no specific CBT tools to support these accommodations.

Exhibit 1.2.3 Accommodations for Students with an Injury

Accommodation	Description
Adult Transcription	<p>An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures.</p> <p>An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.</p>
Assistive Technology	<p>Use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student’s responses exactly as written to the CBT test form or PBT test booklet.</p> <p>Any print copy must be shredded. Any electronic copy must be deleted.</p> <p>This accommodation also requires Adult Transcription.</p>
Rest/Breaks	<p>Student may take breaks during testing sessions to rest.</p>

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. This includes English Language Learner (ELL) students and students withdrawn from English

language services at parent request. Reclassified Fluent English Proficient (FEP) students are monitored for two school years. These FEP Year 1 and FEP Year 2 students also may use, as appropriate, any of the universal test administration conditions and any of the following accommodations.

The *upon student request* accommodations are required to be administered in a setting that does not disturb other students such as in a one-on-one or very small group setting.

Exhibit 1.2.4 summarizes accommodations that may be provided for ELL and FEP students.

Exhibit 1.2.4 Allowable Accommodations for ELL and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request.</p> <p>Reading aloud the content of the Reading portion of the ELA test is prohibited.</p>
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	<p>Exact oral translation, in the student’s native language, of the scripted directions or the directions that students read on their own upon student request.</p> <p>Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted.</p> <p>Translation of test content is not permitted.</p>
Translation Dictionary	<p>Provide a word-for-word published, paper translation dictionary.</p> <p>Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.</p>

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 1.2.5, as designated in their IEP or 504 plan.

Exhibit 1.2.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	<p>An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures.</p> <p>An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.</p>
Assistive Technology	<p>Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted.</p> <p>This accommodation also requires Adult Transcription.</p>
Braille Test Booklet	<p>Provide a paper Braille test booklet.</p> <p>This accommodation also requires Adult Transcription on a regular size paper test booklet.</p>
Large Print Test Booklet	<p>CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet.</p> <p>A PBT Large Print test booklet requires Adult Transcription on a regular size paper test booklet.</p> <p>PBT – Provide a Large Print test booklet.</p> <p>This accommodation also requires Adult Transcription on a regular size paper test booklet.</p>
Paper Test Booklet	<p>CBT – Provide a regular size paper test booklet for a student at a school administering the CBT.</p> <p>If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.</p>

1.3 EVIDENCE BASED ON TEST CONTENT

Because the AzMERIT assessments are designed to measure student progress toward achievement of the ACCRS the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the academic standards.⁶

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards and refers back to those standards in a highly iterative test development process that includes ADE, test developers, and educator committees. Items used to develop the spring 2016 operational test forms were mainly drawn from the AIRCore pool of items developed to align with the Common Core State Standards. These items were also reviewed by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. Only items that were found to align well with the ACCRS were used. To supplement the AIRCore pool of items, a few previously developed Arizona items that also aligned to the ACCRS were used. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning ADE and Arizona educator committees act as reviewers throughout the item development cycle.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards will be covered in each test administration.⁷ Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet the blueprints. Developing test forms is difficult because test blueprints can be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. And in addition to meeting complex blueprint requirements, test developers must work to meet psychometric goals so that alternate test forms measure equivalently across the range of ability.

Following a standard item review process, item reviews proceeded initially through a series of internal reviews before items were eligible for external review by ADE staff and educator committees. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passed through four internal review steps before it was eligible for external review. Those steps include

⁶ Standard 12.4 – When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.

⁷ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

- Preliminary review, conducted by a group of AIR content area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert.

At every stage of the item review process, beginning with preliminary review, AIR's test developers analyze each item to ensure that

- The item is well-aligned with the intended content standard
- The item conforms to the item specifications for the target being assessed
- The item is based on a quality idea (i.e. it assesses something worthwhile in a reasonable way);
- The item is properly aligned to a depth of knowledge (DOK) level;
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward
- Any accompanying graphic and stimulus materials are actually necessary to answer the question
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question
- For selected response items, the set of response options are succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; all plausible, all non-keyed response options are unambiguously incorrect;
- There is no obvious or subtle clueing within the item
- The score points for constructed-response items are clearly defined
- For machine-scored constructed-response (MSCR) items, that item responses yield the intended score points based on the rubric, and
- For human scored constructed response items, the scoring rubric clearly explains what characterizes responses at each possible level of achievement.

In addition, rubric-scored items, both machine-scored and human-scored, are validated following field test administration. Machine-scored items go through a rubric validation process wherein samples of student responses are reviewed, along with resulting scores, to ensure that rubrics are enacted as intended. This process is described in Section 10.1.1. Human-scored items go through a rangefinding process prior to scoring where samples of item responses are used to create scorer training materials and ensure the scoring rubric is appropriate, as described in Section 10.1.2.

Based on their review of each item, the test developer may accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to ADE for their review. At this stage, items may be further revised based on any edits or changes requested by ADE, or rejected outright. Items passing through ADE review level then have to pass through a stakeholder review in which a committee of educators reviews each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- aligned to the content standards

- appropriate for the grade level
- accurate
- presented online in a way that is clear and appropriate
- free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics.

Items were also passed through to a parent/community sensitivity review committee to ensure that test content did not violate community standards. Items successfully passing through both the educator and parent/community review process were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is therefore an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass a three-stage review to be included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct and there are no other obvious problems with the items.

ADE content and psychometric staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, ADE determined that certain flagged items must be rejected, or deemed the item eligible for inclusion in operational test administrations.

1.4 EVIDENCE FOR INTERPRETATION OF PERFORMANCE STANDARDS

Alignment of test content to the Arizona College and Career Ready Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the Arizona College and Career Ready Standards. However, the interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the Arizona standards. AzMERIT test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the Arizona College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Arizona standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standard for the AzMERIT assessments are therefore central to the validity of test score interpretations.⁸

Following the first operational administration of the AzMERIT in spring 2015, a standard setting workshop was conducted to recommend to ADE a set of performance standards for reporting student achievement of the Arizona

⁸ Standard 4.22 – Test developers should specify the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.

College and Career Ready Standards. Arizona educators, serving as standard setting panelists, followed a standardized and rigorous procedure to recommend performance level cut scores. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists used their expert knowledge of the Arizona College and Career Ready Standards and student achievement to map the performance level descriptors adopted by Arizona onto an ordered item book comprising the spring 2015 operational test form and augmented with items administered in the embedded field test slots to minimize information gaps in the operational test form.⁹

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. For each assessment, panelists were provided the approximate location of performance standards for other important assessment systems. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

In addition, panelists were provided with feedback about the vertical articulation of their recommended performance standards so that they could view the relationship between the locations of recommended cut scores for each grade level assessment to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces the interpretation of test scores as indicating not only achievement of current grade level standards, but also preparedness to benefit from instruction in the subsequent grade level.

Following recommendation of final performance standards, the recommended cut scores were presented to the Arizona State Board of Education for review and adoption. The Board adopted the recommended performance standards in August 2015.

Based on the adopted performance standards, Exhibit 1.4.1 shows the estimated percentage of students meeting the AzMERIT proficient standard for each assessment in spring 2015. Exhibit 1.4.1 also shows the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. Exhibit 1.4.1 also presents the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring

⁹ Standard 1.18 – When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

2014 field test administration. As Exhibit 1.4.1 indicates, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college ready, and the NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 1.4.1 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Test	Percent of Students Meeting Standard			
	AzMERIT Proficient	Arizona ACT College Ready	Arizona NAEP Proficient	Projected SBAC
ELA				
3	41%			38%
4	38%		28%	41%
5	30%			44%
6	34%			41%
7	33%			38%
8	32%		28%	41%
9	27%			
10	30%			
11	25%	34%		41%
Math				
3	42%			39%
4	42%		42%	38%
5	40%			33%
6	32%			33%
7	31%			33%
8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

Although AIR previously identified ACT college ready cut scores on the AzMERIT ELA and math scales, that study involved an indirect linkage. In that study, student performance on the grade 10 AIMS was used to predict subsequent student performance on the ACT tests, and then a linking study between the AIMS and AzMERIT allowed for the identification of the ACT cut scores on the AIMS scale to be represented onto the AzMERIT scale.

To examine directly the relationships between the AzMERIT and ACT assessments, ADE obtained the ACT test scores for Arizona students graduating high school in spring 2016. Although AzMERIT is offered as a series of end of course tests in high school, most students take the Algebra II assessment at grade 11, so the focus of this investigation will be on the grade 11 ELA and Algebra II AzMERIT assessments administered in spring 2015.

Because a selected sample of fewer than half of Arizona students takes the ACT, a two-step approach is typically adopted to impute missing data in the analysis of the relationship between the AzMERIT and ACT test scores. However, previous investigations with other state assessments, including the Arizona AIMS, demonstrated that imputing or deleting the missing records did not impact the linkage identified between state assessments and the ACT test. For this study we instead divided the complete sample of merged records into model building and cross-validation samples of equal size. The cross-validation sample allows for better estimation model fit. Because the

model is built using a sample independent from that used to evaluate model fit, estimates of model fit exclude sample dependent idiosyncrasies that would be reflected as model overfit in the model development sample.

Exhibit 1.4.2 shows the location of the ACT college ready cut scores for math and reading on the AzMERIT scale. The first column shows the location as identified via indirect linkage through AIMS and that was provided as benchmark information to AzMERIT standard setting panelists. The second column shows the location of the ACT college ready cut scores as identified via direct linkage between ACT and AzMERIT described here. The third column shows the location of the AzMERIT meets performance standard on the Algebra II and Grade 11 ELA assessments. As indicated in the table, the location of the ACT college ready cut scores on the AzMERIT scale were reasonably consistent across methods, especially for ELA. Importantly, the results affirm that the location of adopted AzMERIT performance standards are consistent with the ACT college ready criteria.

Exhibit 1.4.2. Location of the ACT College Ready Cut Scores on the AzMERIT Scales.

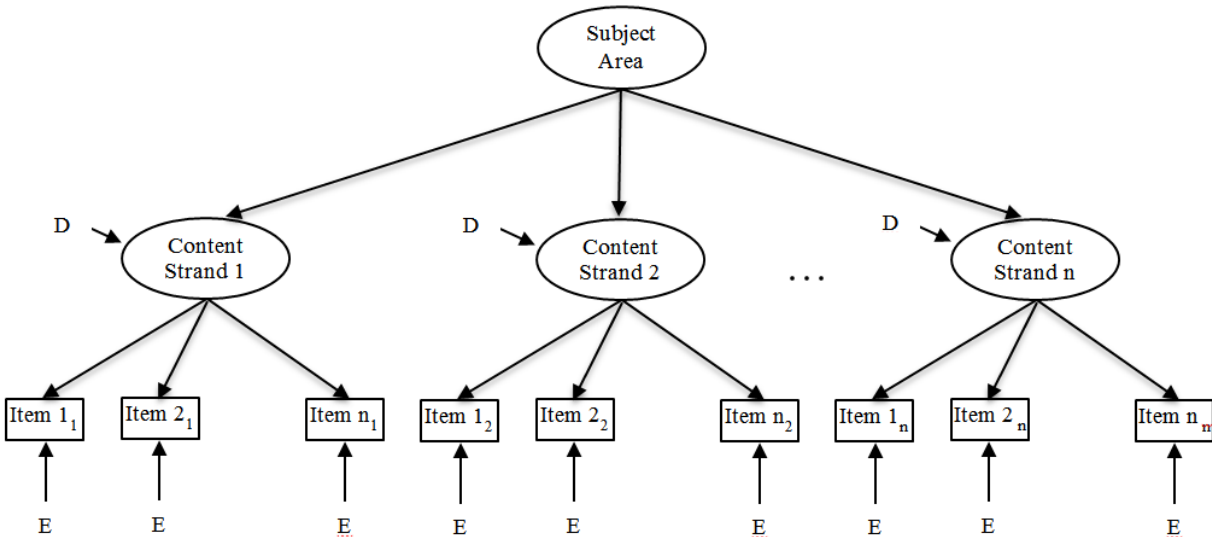
	Location of ACT College Ready Cut on AzMERIT Scale		AzMERIT Meets Performance Standard
	Via Indirect Linkage through AIMS	Via Direct Linkage with AzMERIT	
Algebra II	3704	3727	3711
Grade 11 ELA	2579	2585	2585

The equipercentile equating method was used to verify the linkage between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT college ready cut scores in reading was 2586 on the AzMERIT ELA scale. The location of the ACT college ready cut score in math was 3727 for the AzMERIT math scale. Results from the equipercentile approach were thus consistent with the cut scores identified using regression models.

1.5 EVIDENCE BASED ON INTERNAL STRUCTURE

Arizona's AZMERIT assessment represents a structural model of student achievement in grade level and course specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., Reading Information, Reading Literature, Language, Writing). Content strands within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Exhibit 1.5.1. As the exhibit illustrates, each item is an indicator of an academic content strand. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content strand serves as an indicator of achievement in a subject area. As at the item level, the content strands include an error term indicating that the content strands are not pure indicators of overall achievement in the subject area. The paths from the content strands to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content strand construct. Similarly, the paths from subject area achievement to the content strands represent the second-order factor loading, indicating the degree to which academic content strand constructs are correlated with the underlying construct of subject area achievement.

Exhibit 1.5.1 Second-Order Structural Model for AzMERIT Assessments



Following the first operational test administration in spring 2015, confirmatory factor analysis was used to evaluate the fit of this structural model to student response data.¹⁰ For each of test forms administered in spring 2015, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size, however, so even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices, such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (Tucker & Lewis, 1973), and the Root Mean Square of Approximation (RMSEA) were also used to evaluate model fit.

The AzMERIT assessments also claim to measure subject area achievement using test items that probe student knowledge and skills across multiple depth of knowledge levels. As with the content standards, the classification of items by depth of knowledge also represents a structural model that can be evaluated using confirmatory factor analysis.¹¹ In this case, each item is an indicator of a depth of knowledge level first-order factor, and each depth of knowledge level is in turn an indicator of subject area achievement. Thus, confirmatory factor analysis was used to evaluate the fit of this depth of knowledge structural model to student response data from the spring 2015 AzMERIT test administrations.

Exhibit 1.5.2 Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Index	Indication of Good Fit
CFI	$\geq .95$
TLI	$\geq .95$
RMSEA	$\leq .05$

¹⁰ Standard 1.13 – If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.

¹¹ Standard 1.12 – If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

In addition to testing the fit of the hypothesized AzMERIT second-order confirmatory factor analysis model, we examined the degree to which the second-order model improved fit over the more general one-factor model of academic achievement in each subject area. Because the second-order model was nested within the one-factor, general achievement model, a simple likelihood ratio test was used to determine whether the added information provided by the structure of the ACCRS frameworks improved model fit over a general achievement model. Results indicating improved model fit for the second-order factor model provide support for the interpretation of content standard performance above that provided by the overall subject area score.¹²

1.5.1 ELA CONTENT MODEL

We began by evaluating the fit of the first-order, general achievement model in which all items are indicators of a common subject area factor. This model importantly evaluates the assumption of unidimensionality of the subject area assessments, and provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the first-order, general achievement models in ELA are shown in Exhibit 1.5.1.1. All of the statistics indicate the general achievement factor model fit the data well. This pattern was true across all grades. The CFI and TLI values were all greater than 0.9 and generally equal to or greater than 0.95, and the RMSEA values were all below .05, indicating good fit for the base model.

Exhibit 1.5.1.1 Goodness-of-Fit for the AzMERIT ELA First-Order Model

Grade	CFI	TLI	RMSEA
3	0.934	0.931	0.047
4	0.949	0.946	0.033
5	0.966	0.964	0.039
6	0.955	0.953	0.043
7	0.974	0.972	0.037
8	0.964	0.963	0.048
9	0.924	0.921	0.039
10	0.948	0.945	0.042
11	0.928	0.925	0.034

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.1.2. All of the statistics indicate the second-order models posited by the AzMERIT assessments fit the data well. This pattern was true across all grades. As with the general factor model, the CFI and TLI values for the second-order models were all equal to or greater than .95, with RMSEA values well below the .05 threshold used to indicate good fit.

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.5.1.3. We note that model fit for first-order model of general achievement are also very high and provide evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether the posited second-order reporting model adds information beyond that provided by the first-order model. The chi-square difference test shows that across grade levels, the strand-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

¹² Standard 1.14 – When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Exhibit 1.5.1.2 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.958	0.956	0.038
4	0.970	0.969	0.025
5	0.980	0.979	0.030
6	0.973	0.972	0.033
7	0.983	0.982	0.029
8	0.980	0.979	0.036
9	0.962	0.960	0.028
10	0.972	0.970	0.031
11	0.949	0.947	0.029

Exhibit 1.5.1.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	df	p value
3	13560.7	3	$p < .001$
4	8460.9	3	$p < .001$
5	10944.7	3	$p < .001$
6	12019.8	3	$p < .001$
7	8848.6	3	$p < .001$
8	15590.1	3	$p < .001$
9	8896.6	3	$p < .001$
10	9084.7	3	$p < .001$
11	4412.8	3	$p < .001$

1.5.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.2.1. Across all grades, results indicate the second-order models posited by the AzMERIT assessments fit the data well. The CFI and TLI values were all .97 to .99, RMSEA values are all .03 or lower.

Exhibit 1.5.2.1 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.99	0.99	0.02
6	0.98	0.98	0.03
7	0.99	0.99	0.02
8	0.99	0.99	0.02
9	0.98	0.98	0.02
10	0.98	0.97	0.02
11	0.98	0.98	0.02

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.5.2.2. The chi-square difference test shows that across grade levels, the DOK-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} p -values were less than .001 across all grade levels.

Exhibit 1.5.2.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	df	p value
3	21402.6	4	$p < .001$
4	12053.6	4	$p < .001$
5	17102.9	4	$p < .001$
6	18192.1	4	$p < .001$
7	16351.4	4	$p < .001$
8	25454.7	4	$p < .001$
9	14989.3	4	$p < .001$
10	14920.9	4	$p < .001$
11	8075.1	4	$p < .001$

1.5.3 MATH CONTENT MODEL

As with ELA, structural analyses of the math assessments began with an evaluation of fit for the first-order, general achievement model in which all items are indicators of a common math subject area factor. This model provides for an evaluation of the unidimensionality assumption of the subject area assessments, and provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the general achievement models in math are shown in Exhibit 1.5.3.1. All of the statistics indicate the general achievement factor model fit the data well. This pattern was true across all grades. The CFI and TLI values were all equal to or greater than .95, and the RMSEA values are all below .05, indicating good fit for the base model.

Exhibit 1.5.3.1 Goodness-of-Fit for the AzMERIT Math First-Order Model

grade	CFI	TLI	RMSEA
3	0.975	0.973	0.027
4	0.976	0.975	0.024
5	0.976	0.975	0.026
6	0.975	0.973	0.023
7	0.982	0.981	0.021
8	0.969	0.967	0.026
Algebra I	0.976	0.975	0.023
Algebra II	0.973	0.971	0.021
Geometry	0.986	0.985	0.018

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 1.5.3.2 (below). The models show very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their .05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.5.3.2 Goodness-of-Fit for the AzMERIT Math Second-Order Model

grade	CFI	TLI	RMSEA
3	0.979	0.978	0.024
4	0.978	0.977	0.024
5	0.978	0.977	0.025
6	0.976	0.975	0.023
7	0.983	0.982	0.021
8	0.970	0.969	0.026
Algebra I	0.978	0.977	0.022
Algebra II	0.974	0.972	0.020
Geometry	0.987	0.986	0.017

The results of the comparison between the second-order, strand-based model and the first-order, general achievement model are presented in Exhibit 1.5.3.3. Again, model fit for the general achievement first-order model is very high, providing evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether knowledge of the depth of knowledge level of items provides information beyond that provided by the more general model. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.3.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	df	<i>p</i> value
3	3225.0	3	<i>p</i> < .001
4	1326.3	3	<i>p</i> < .001
5	1427.0	3	<i>p</i> < .001
6	1036.2	4	<i>p</i> < .001
7	559.8	4	<i>p</i> < .001
8	1039.3	4	<i>p</i> < .001
Algebra I	750.9	3	<i>p</i> < .001
Algebra II	246.5	3	<i>p</i> < .001
Geometry	269.7	4	<i>p</i> < .001

1.5.4 MATH DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 1.5.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their .05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.5.4.1 Goodness-of-Fit for the AzMERIT Math Second-Order Model

grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.99	0.99	0.02
Geometry	0.97	0.97	0.02

The results of the comparison between the second-order, DOK-based model and the first-order, general achievement model are presented in Exhibit 1.5.4.2. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.4.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	df	p value
3	331.4	3	$p < .001$
4	309.5	3	$p < .001$
5	14.9	3	$p < .001$
6	14.5	3	$p < .001$
7	236.6	3	$p < .001$
8	79.2	3	$p < .001$
Algebra I	20.1	3	$p < .001$
Algebra II	26.4	3	$p < .001$
Geometry	20.9	3	$p < .001$

1.6 EVIDENCE FOR RELATIONSHIPS WITH CONCEPTUALLY RELATED CONSTRUCTS

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity. Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct irrelevant attributes.¹³

Observed correlations between alternate indicators of student achievement of course objectives, such as locally administered assessments of student achievement and AzMERIT, should be limited only by the unreliability of the measures. When both assessments measure student achievement in common subject areas, as with for example, locally administered and statewide assessments of math achievement, we expect test scores between the common subject area assessments to be substantially correlated. In addition, we expect that the magnitude of observed correlations between test scores in different subject areas will be lower than correlations between test scores in a common subject area. Because the content domains assessed in ELA and math tests are quite different, AzMERIT ELA test scores should correlate less well with locally administered assessments of math than ELA. It is important to note, however, that test scores across subject areas and test systems are nevertheless expected to be highly correlated. This is because even though subject area test scores measure different academic content domains, student achievement across subject areas is influenced by factors both internal (e.g., general intelligence) and external (e.g., socioeconomic status) to the student that contribute to student achievement across all academic subject areas so that student test scores across subject areas tend to be highly intercorrelated. So while we certainly do expect correlations between test scores across subject areas to be lower than correlations between test scores within a subject area, we nevertheless expect test scores across subject areas to be quite high.

Exhibit 1.6.1 shows the correlations between student test scores on the spring 2015 statewide AzMERIT assessment with corresponding test scores on a district-wide administration of the Northwest Evaluation Association (NWEA) assessment. Sample sizes range from more than 1,400 students taking the grade 3 assessments, to nearly 1,100 students taking the middle school assessments, so the observed correlations are expected to be stable. Convergent

¹³ Standard 1.16 – When validity evidence includes empirical analyses of responses to test items together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

correlations are quite high, ranging from 0.82 to 0.84 between AzMERIT ELA (assessing reading, writing, and listening) and NWEA reading. Correlations between AzMERIT and NWEA math scores are even higher, ranging from 0.85 to 0.89.

Exhibit 1.6.1 Correlations between AzMERIT and Locally Administered NWEA Test Scores

Grade	ELA Sample Size	ELA Correlation	Math Sample Size	Math Correlation
3	1426	0.82	1429	0.86
4	1214	0.84	1214	0.88
5	1303	0.84	1303	0.88
6	1119	0.82	1115	0.85
7	1081	0.82	1082	0.89
8	1090	0.82	1091	0.89

Exhibit 1.6.2 shows the discriminant correlations between AzMERIT and the locally administered NWEA assessment. As expected, correlations across subject area assessments remain quite high, indicating considerable consistency in student achievement across subject area assessments. Nevertheless, correlations across subject area assessments are systematically lower than within subject correlations, indicating that the subject area assessments are measuring domain specific knowledge and skills in addition to common factors underlying student achievement.

Exhibit 1.6.2 Discriminant Correlations between AzMERIT and Locally Administered NWEA Test Scores

Grade	ELA Sample Size	ELA Correlation	Math Sample Size	Math Correlation
3	1426	0.72	1428	0.70
4	1211	0.76	1217	0.72
5	1303	0.75	1303	0.72
6	1117	0.73	1117	0.71
7	1081	0.77	1080	0.74
8	1088	0.75	1093	0.71

Convergent correlations between AzMERIT and locally administered assessments were also reported by Estrada and colleagues (Estrada, Burnham, Feld, Bergan, and Bergan, 2015). These researchers reported the mean correlations between a variety of local assessments and AzMERIT test scores for ELA and math assessments in grades 3-8. Mean correlations between AzMERIT and various local assessments of ELA ranged from .77 to .79 across the grade levels investigated. Mean correlations between AzMERIT and local assessments of mathematics ranged from .71 to .75 across grade levels 3 through 8. These results likewise show good convergence between AzMERIT and other locally administered assessments purporting to measure the same constructs.

1.7 MEASUREMENT INVARIANCE ACROSS SUBGROUPS

Measurement invariance occurs when the likelihood of responding correctly conforms to the measurement model and is independent of group membership and the parameters of a measurement model are statistically equivalent across groups.¹⁴ The parameters of interest in measurement invariance testing are the factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that

¹⁴ Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

it is not necessary to demonstrate invariance across groups on these parameters. In general, measurement invariance testing can be conducted using a series of multiple-group confirmatory factor analysis (CFA) models, which impose identical parameters across groups. The measurement model parameters, including factor patterns (configural invariance), factor loadings (metric or weak invariance), latent intercepts/thresholds (scalar or strong invariance), and unique or residual factor variances (strict invariance), are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups.

Appendix B shows the results of measurement invariance testing by subgroups for ELA and Math. Items comprising the spring 2016 operational test administration were used to investigate measurement invariance across subgroups. The full set of tables associated with these analyses is provided for each of the grade and subject area assessments. The series “a” tables (e.g., tables B.1a, B.2a, etc.) present the global model fit indices for the measurement invariance tests for each assessment. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using χ^2 difference test (at $\alpha \leq 0.05$) and the examination of significant differences of the Root Mean Square of Approximation (RMSEA, change in RMSEA ≤ 0.015 ; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across the following subgroups: gender (Model A), ethnicity including African American vs. White (Model B-1), Hispanics vs. White (Model B-2), Asian vs. White (Model B-3), American Indian vs. White (Model B-4), and Multi-Ethnics vs. White (Model B-5), special education program status (SPED; Model C), economic disadvantage status (Low Income; Model D), limited English proficiency status (LEP; Model E), and accommodated test forms (Accommodation, Model F). Invariance tests of subgroups were investigated separately for each grade and subject area test. Since in each ELA assessment, students were randomly assigned to one of six writing prompts for administration, the missing responses on the writing items resulted in unsuccessful model convergence. Thus, to achieve model convergence, we included the students who took a common writing prompt between online and paper in each ELA assessment.

The null hypothesis of the χ^2 difference test is that the more restricted invariance model (e.g., metric) fits the data equally as well as the less restricted invariance model (e.g., configural). Given that the sensitivity of the χ^2 difference tests to sample size, we additionally examined significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retention of the more restricted invariance model (Chen, 2007).

The series “b” tables (e.g., tables B.1b, B.2b, etc.) show the model fit indices of scalar invariance models assuming same factor pattern + identical factor loadings + identical latent intercept/threshold across subgroups. Global model fit indices included the Comparative Fit Index (CFI; Bentler, 1990) and Root Mean Square of Approximation (RMSEA). CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.870 to 0.990 and RMSEA ranged from 0.012 to 0.044. For Math, CFI values ranged from 0.905 to 0.990 and RMSEA ranged from 0.010 to 0.058.

Although the χ^2 difference test should ideally be nonsignificant, all χ^2 difference tests were significant at $\alpha = .05$ due to large sample sizes except Model B-5 where the χ^2 difference tests for most grades was nonsignificant or marginally significant at $\alpha = .05$. In spite of significant χ^2 difference tests for most models, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0.000 to 0.002 for both ELA and MATH). Based on the similar magnitudes of the RMSEA (i.e., no material change across all tested models; Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, ELA and MATH test scores have the same measurement structure across gender, ethnicity (African American vs. White, Hispanics vs. White, Asian vs. White, American Indian vs. White, and Multi-Ethnics vs. White), special education program status, economic disadvantage status, limited English proficiency status, and accommodation test forms.

1.8 DIFFERENTIAL MODE EFFECTS ACROSS SUBGROUPS

To explore the possibility that mode of test administration may exert differential effects across subgroups, we began by identifying matched samples of students participating online and on paper. For students administered assessments on paper, observed test scores were regressed on prior achievement and demographic variables to obtain regression weights. The resulting prediction equation was then applied to all students to yield predicted paper test scores. The predicted paper scores were used to identify matched samples of online and paper test takers.

To identify possible differential effects of mode across subgroups, we used the observed test score as dependent variable and then covaried the predicted test score to isolate the effects of mode. We then entered the dummy-coded demographic variables including gender, English language learner status, special education status (SPED), free reduced lunch status (FRL), migrant status, and six ethnicity subgroups as predictors. Significant interactions between mode of test administration and the demographic subgroup comparisons indicate differential mode effects between the specified demographic subgroups.

While many effects achieve conventional levels of statistical significance, because of the very large sample sizes, the effect sizes were quite small. Thus, Exhibit 1.8.1 shows the regression coefficient estimates for the differential mode effects by subgroup interaction only for effects where $p < .0001$.

Results indicated that mode effects were more pronounced for special education students relative to general education population. Especially for the high school EOC tests, AzMERIT tests were more difficult for special education students when administered on paper than online.

Mode effects were more pronounced for low income students with respect to the math assessments. Math tests were generally more difficult for low income students when administered online than on paper.

Mode effects were also more pronounced for LEP students than for the general education population in math but not ELA. However, the direction of this effect was not consistent across grades. Online math tests were more difficult than paper for LEP students in the lower grades, while paper math tests were more difficult than online tests for LEP students in the higher grades.

Exhibit 1.8.1 Parameter Estimates for Differential Mode Effects by Subgroups Interactions

Test	Gender	White	Black	Asian	Hawaii/Pacific	Hispanic/Latino	American Indian	Special Education	Limited English Proficiency	Free/Reduced Lunch	Migrant
ELA											
G3E	0.49									0.27	
G4E											
G5E											
G6E								-0.61			
G7E								0.5			
G8E					1.66	-0.34					
G9E	0.45							-0.74			
G10E								-1.23		-0.41	
G11E	-0.33					0.36		-0.58			
MATH											
G3M								0.57			
G4M									0.52	-	-4.46
G5M							-0.89			0.34	
G6M		1.15	0.96				0.69		0.6	-0.31	
G7M	-0.26									0.25	-2.87
G8M		0.89					0.86		-0.58		
Algebra I						0.73		-0.8	-0.95	0.5	
Geometry						-0.44		-1.32		1.11	
Algebra II							-1.07	-0.75		0.63	

Note: Positive coefficient means that the online test is more difficult for the focal group.

1.9 EVIDENCE FOR STUDENT GROWTH – OVERALL AND BY SUBGROUPS

The AzMERIT assessments report student test scores on a vertical scale, allowing families and teachers to make inferences about student growth across school years. The validity of test score interpretations about student growth over time depend strongly on the vertical linking design used to develop the vertical scale. But even when test score interpretations are appropriate to the scaling design, it is important to examine whether student gains may be interpreted consistently across subgroups or whether differential gain rates across subgroups limit the inferences that can be made about test score gains over time.¹⁵ To address this issue, we examined rates of

¹⁵ Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Standard 3.17 – When aggregate scores are publicly reported for relevant subgroups— for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults— test users are responsible for providing evidence of comparability and

student growth across student gender, race/ethnicity, students with disabilities (SPED), English language learners (LEP), and low income status (Low Income).

Exhibit 1.9.1 shows the mean test scores on the spring 2015 and spring 2016 administrations of AzMERIT for students participating in both test administrations, as well as the correlation between test scores across the two assessment occasions. Correlations between test scores are quite high and indicate substantial consistency in rank ordering of student achievement between the two test administrations. The correlation between student achievement in grade 8 math and Algebra I is attenuated somewhat, and further that the distribution of student ability is somewhat less variable for this cohort, especially with respect to the spring 2016 Algebra I performance. We note that in spring 2015, grade 8 students enrolled in Algebra I were required to participate in both assessments, but in spring 2016, those high achieving students would likely have participated in the Geometry assessment and would not have been included in these analyses. The resulting restriction of range could be responsible for the attenuated correlation.

The exhibit also shows that rate of achievement gain is somewhat higher for math than ELA, and that while gain rates decelerate across the school years, the rate of gains diminishes more rapidly for ELA than math over time. For math, large gains, typically $\frac{3}{4}$ standard deviation or more (e.g., average gain of 33 scale score points in grade 3 math is 80% of the 40 point standard deviation of student test scores), are observed through the middle school grades, dropping to about $\frac{1}{3}$ standard deviation between administrations of the high school end-of-course assessments. For ELA, while elementary school gains are strong, by middle school, annual gains are between $\frac{1}{3}$ to $\frac{1}{2}$ standard deviation, and by high school drop to about $\frac{1}{4}$ standard deviation, with no growth observed between grade 10 and 11.

Exhibit 1.9.1 Test Score Stability and Performance Gains Overall

Assessment Subject_2015_2016	N	Spring 2015 Scale Score		Spring 2016 Scale Score		Change from 2015 to 2016		Percent Scoring Lower		Correlation
		Mean	Std Dev	Mean	Std Dev	Mean	IRT based Standard Error	Expected	Observed	
ELA										
ELA_G3E_G4E	80245	2501	29.99	2518	33.59	17	14.43	0.23	0.18	0.82
ELA_G4E_G5E	79662	2514	28.61	2537	34.08	23	14.32	0.16	0.11	0.82
ELA_G5E_G6E	78965	2529	27.9	2541	33.59	13	14.06	0.29	0.24	0.83
ELA_G6E_G7E	78273	2541	29.81	2553	30.88	11	13.90	0.30	0.25	0.84
ELA_G7E_G8E	76782	2548	28.85	2556	32.21	8	13.74	0.35	0.32	0.84
ELA_G8E_G9E	69223	2561	28.67	2566	30.81	6	13.72	0.40	0.36	0.82
ELA_G9E_G10E	61972	2562	26.31	2567	28.68	5	13.54	0.41	0.38	0.80
ELA_G10E_G11E	53924	2571	26.48	2569	31.19	-2	13.31	0.53	0.54	0.80
MATH										
Math_G3M_G4M	80875	3522	39.27	3553	40.21	31	16.38	0.14	0.10	0.81
Math_G4M_G5M	80277	3553	37.49	3589	41.29	36	16.11	0.11	0.07	0.81
Math_G5M_G6M	79107	3585	38	3616	42.37	31	16.05	0.14	0.09	0.83

for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Math_G6M_G7M	76152	3614	34.1	3633	35.48	20	15.22	0.21	0.16	0.83
Math_G7M_G8M	64887	3625	33.3	3652	35.99	27	15.48	0.14	0.09	0.82
Math_G8M_Algebra I	50295	3651	29.23	3665	28.79	13	14.98	0.30	0.25	0.74
Algebra I_Geometry	54203	3674	33.74	3685	35.71	11	15.72	0.33	0.29	0.80
Geometry_Algebra II	43737	3687	33.22	3697	33.17	11	16.37	0.34	0.31	0.77

To evaluate differential growth across demographic subgroups, a series of regression analyses were conducted to predict 2016 test scores from 2015 test scores, controlling for demographic subgroup membership. To compare ethnic subgroup performance, we created six dummy variables contrasting white students with each of other ethnic groups (e.g., white/Hispanic, white/African American). Gender was coded 1 for female. SPED, LEP, and Low Income students were coded as 1 to contrast with students who were not identified with those needs who were coded as 0.

Appendix C shows the regression model parameter estimates for the ELA and math assessments. The 2015 test scores were centered on the reference group mean so that the intercept values at the top of the table represent the mean performance of white males on the 2015 assessment, with group parameters reflecting differences from the reference group on the spring 2016 assessment. Results indicate that females generally performed better than males for both ELA and math across grades. With respect to ethnicity, Asian students generally performed better than white students in both ELA and math. For all other ethnic group comparisons, the focal groups generally performed less well than whites. Special education students, limited English proficient students, and low income students all performed less well than the general education population in both ELA and math.

The slope represents the association between 2015 and 2016 test scores, controlling for demographic subgroups. The overall positive slope parameter indicates reference group gains in test scores between 2015 and 2016. The group specific slope parameters indicate differential gains between contrasted groups.

Although many individual effects attained conventional levels of statistical significance due to large sample sizes, we focus here only on highly significant effects that are associated with more practically significant effect sizes and that may point to trends across grade level and/or subject area assessments.

While females tended to score higher across assessments, differential gain rates by gender were small and inconsistent.

With respect to ethnicity, differential gain rates were small and inconsistent in the elementary and middle school grade assessments. Asian students did, however, show higher gain rates than whites only in the high school years. And African American and Hispanic students showed lower gain rates than whites in the high school math assessments.

Special education students generally showed lower rates of gain than general education students, although pattern was reversed for elementary school math, with special education students showing greater rates of gain from grade 3 to 4, and from grade 4 to 5.

Limited English proficient students showed lower rates of gain in both ELA and math, but this effect seems to moderate in the high school grades where differential gain rates were much less pronounced.

Differential gain rates for low income students were observed for ELA, but were inconsistent across grades. Low income students generally showed lower gain rates in math, but this effect was more apparent at the middle and high school grades.

1.10 DAY, WEEK, AND TIME OF DAY EFFECTS ON PERFORMANCE

Administration of Arizona's new AzMERIT online tests is untimed and schools may flexibly schedule students to take the tests in computer labs throughout the testing window. Thus, students taking the same grade level or end of course (EOC) test are not required to test on the same day. This is a marked departure from Arizona's administration of AIMS in which students statewide were administered tests on the same date and nearly the same time, ensuring that any effects of test administration time or day were held constant.

Because the days and times on which tests can be administered is variable, the possibility arises that performance factors associated with time of day or day of week may influence student test scores.

A series of regression models were developed to predict student performance using the day of the week and time of the day variables, as well as the duration of the test administration from test start to test end. The dependent variable for these analyses was the spring 2016 AzMERIT scale score. To control for student achievement, we first covaried previous achievement using spring 2015 AzMERIT test scores. Because of the need to covary previous achievement, the analyses were limited to students participating in the grade 4 to 8 and high school EOC assessments in mathematics and ELA tests, and for whom 2015 test scores were available. The day of the week was coded as 1 to 5 (1 for Monday, 2 for Tuesday, and so on). For the regression analyses, the time of day and the duration were continuous variables using the actual time. Time of day effects were further evaluated using paired comparisons between early morning, late morning, early afternoon, and late afternoon.

Exhibit 1.10.1 shows the standardized regression coefficient estimates of the time effect on student's performance only for effects where $p < .05$. Results indicate generally that starting tests earlier in the week resulted in higher test performance. Tests started on Friday were consistently associated with impaired performance. There were some exceptions to this. For example, students beginning the grade 7 ELA tests on Monday scored lower than students beginning on any other day than Friday. But generally the pattern was pronounced.

Conversely, assessments which were completed earlier in the week were associated with lower test scores. Tests ending on any other day than Monday were associated with higher test scores. And this effect was generally true for tests ending on Tuesday. That said, students appeared to perform better on tests ending Wednesday or Thursday than on Friday, although there were exceptions to this as well (e.g., grade 9 and grade 10 ELA where Friday end dates were associated with greater performance).

Time of day effects were less consistent. For ELA, morning start times were associated with greater performance than afternoon start times for high school students. For middle school students later morning start times were associated with poorer performance than early morning or late afternoon. And at grade 5, ELA tests with morning start times were associated with lower performance than tests with afternoon start times.

For math tests, later start times were generally associated with better performance. An exception to this pattern was observed for Algebra I, where students beginning testing late morning performed better than students starting at any other time.

Tests ending early in the afternoon were generally associated with higher performance than tests ending earlier in the day, although grade 6 ELA proved an exception with tests ending early morning associated with the highest scores.

In addition, longer test administrations were associated with higher performance.

Exhibit 1.10.1. Standardized Regression Coefficients of Time Effect on Student's Performance

Test	Start Day	End Day	Start Time	End Time	Duration
ELA					
Grade 4 ELA		0.02	-0.01	0.03	-0.01
Grade 5 ELA	-0.01	0.01	-0.01	0.02	
Grade 6 ELA	0.02		0.01		
Grade 7 ELA	0.01	0.03	-0.01	-0.01	0.01
Grade 8 ELA		0.02	-0.01		0.02
Grade 9 ELA		0.01	-0.06	0.02	0.01
Grade 10 ELA	-0.02		-0.08	0.03	0.01
Grade 11 ELA	-0.03		-0.08	0.05	0.01
MATH					
Test	Start Day	End Day	Start Time	End Time	Duration
Grade 4 MATH	-0.01	0.02	-0.02		
Grade 5 MATH	-0.02	0.01	-0.03	0.04	0.01
Grade 6 MATH	-0.03	0.01		0.03	0.01
Grade 7 MATH	-0.01	0.01	-0.04	0.06	
Grade 8 MATH		0.01	-0.01	0.04	
Algebra I	-0.05	0.01	-0.12	0.08	0.04
Geometry		0.03	-0.11	0.10	0.03
Algebra II	-0.04	0.04	-0.13	0.12	0.05

Note: Standardized regression coefficient 0.01 is equivalent to 3 or 4 scale score difference.

1.11 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretation is ongoing. Nevertheless, there currently exists sufficient evidence to support the principle claims for the test scores, including that AzMERIT test scores indicate the degree to which students have achieved the Arizona College and Career Ready Standards at each grade level, and that students scoring at the proficient level or higher demonstrate levels of achievement consistent with national benchmarks indicating that they are on track to college readiness. These claims are supported by evidence of a test development process that ensures alignment of test content to the ACCRS, a standard setting process that yielded performance standards consistent with those of rigorous, national benchmarks. Confirmatory factor analyses indicate that the subject area assessments are unidimensional and therefore consistent with the measurement model, but also that the hypothesized reporting strand structure of the AzMERIT provides significant additional information about student achievement. In addition, test scores on the AzMERIT correlate strongly with other measures of subject area achievement, and demonstrate differential relationships across subject area assessments.

2. BACKGROUND OF ARIZONA STATEWIDE ASSESSMENTS

In November 2014, the Arizona State Board of Education adopted Arizona's Measurement of Educational Readiness to Inform Teaching, or AzMERIT, to measure student mastery of the Arizona academic standards and progress toward college and career readiness. The AzMERIT measures English language arts in grades 3-11, and math in grades 3-8 and following completion of high school coursework in Algebra I, Geometry, and Algebra II. ADE worked with the American Institutes for Research to develop and administer the AzMERIT beginning in the spring of 2015. In accordance with state requirements, the AzMERIT was designed to¹⁶:

- Align to the academic standards adopted by the Arizona State Board of Education in 2010 (Arizona College and Career Ready Standards, or ACCRS)
- Supply criterion referenced summative assessments for grades 3 through 8, and criterion referenced end of course assessments in identified high school math and English language arts courses for implementation beginning in the 2014-15 school year
- Assess, without bias, a range of basic knowledge and lower level cognitive skills and higher order, analytical thinking skills in writing, analysis, and problem-solving across subjects, using multiple assessment methods
- Provide valid, reliable and timely data to educators and policy makers to advance the academic success of Arizona students and inform the State's accountability measures
- Communicate results to students, parents and educators, in a clear and timely manner to guide instruction
- Provide an accurate perspective of the quality of learning occurring within classrooms and schools
- Offer educators, students, and families critical tools to improve student achievement, including, but not limited to, formative and interim assessments, sample items and practice tests
- Allow meaningful national or multistate comparisons of school and student achievement
- Use 21st Century technology to deliver the assessment, as available infrastructure allows
- Ensure clarity, transparency, accuracy and security in all aspects of assessment development, deployment, scoring and reporting
- Provide for content and psychometric evaluation and validation
- Establish the involvement of Arizona stakeholders – educators, students, parents, institutions of higher education, and business – in the development of the test, test related materials, and achievement levels indicative of college and career readiness
- Demonstrate accessibility for all students, with optimal access for English language learners and students with special needs
- Respect Arizona's local control of the selection of classroom instructional materials
- Satisfy assessment goals in a cost-efficient manner

The AzMERIT was first administered in spring 2015, assessing proficiency in ELA in grades 3 through 11, math in grades 3 through 8, and following completion of Algebra I, Geometry, and Algebra II (or similar) coursework.

¹⁶ Standard 7.1 – The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

Following the initial administration, the AzMERIT in grades 3 through 8 will be administered in the spring of each academic year; tests assessing high school end-of-course tests will be administered in the fall, spring, and summer of each academic year.

The Rasch model, and Masters' (1982) Partial Credit Model, an extension of the one parameter Rasch model that allows for graded responses, was used to estimate item parameters for the AzMERIT. Item pools for grade level summative and End of Course assessments were calibrated following the first operational administration in spring 2015. A vertical linking design was also implemented to produce a common vertical scale across grade levels to monitor student growth across grades 3 through 8, as well as the high school EOC assessments. In subsequent years, pre-equated bank item parameter estimates will be applied directly for final scoring and reporting, a strategy that allows for more rapid reporting of tests administered online.

2.1 DEVELOPMENT OF ARIZONA COLLEGE AND CAREER READY STANDARDS (ACCRS)

In 2010, the Arizona State Board of Education adopted new academic content standards in ELA and mathematics that reflect high expectations all Arizona students and strive to ensure that high school graduates are college and career ready. The Arizona College and Career Ready Standards (ACCRS) in mathematics describe expectations for learning in grades K-8 and the first three high school courses (Algebra I, Geometry, Algebra II, or Mathematics 1,2,3) plus specific standards that could be included in a fourth high school credit mathematics course. The ACCRS in ELA describe the reading, writing, language, and speaking and listening skills that students should acquire from grade K-12. The standards can be found [on ADE's website](#).

2.2 AZMERIT TEST DESIGN

The AzMERIT is a series of fixed form assessments that are intended to be administered online, although the assessment is offered as a dual mode, online and paper, assessment to accommodate schools that are not yet ready to transition to the online testing environment. A common operational base form is administered to all students within a given test grade and subject. Each assessment is comprised of two to three discrete test sessions. The AzMERIT operational item pools include a variety of selected response, machine-scored constructed responses (MSCR), and some hand-scored constructed-response items in the paper math forms where MSCR items could not readily be rendered for paper test administration. AzMERIT also includes essay responses. In spring 2016, a sample of online responses was hand-scored (100% double scoring with resolution of all discrepancies) for purposes of developing statistical models for machine scoring the remaining online responses.

Six types of MSCR items were included in the AzMERIT forms: graphic response, natural language, equation response, hot text, and table input items. The graphic response item types require students to place objects or move objects around in the answer space. A student can also plot points, draw lines, and draw shapes. The natural language item types require students to type an English language answer. The equation response items require students to enter a value or equation. Hot text items ask students to select or rearrange sentences or phrases in a passage. The table input item types require students to input numerical values into a table. The validity of computer-assigned scores for constructed-response items was evaluated following the spring 2015 online administration of the embedded field test items. Rubric validation for all operational test items was completed prior to test construction and was based on the previous field test administration of those items.

Each ELA assessment included one writing essay prompt that required an extended essay response. For the online test administrations, students were randomly administered one of six writing tasks. A random sample of student responses to each writing task was selected for human scoring. These responses were scored by two human raters

on three distinct scoring dimensions or rubrics: Statement of Purpose/Focus and Organization, Evidence/Elaboration, and Conventions/Editing, with any discrepancy adjudicated in a resolution score. This sample of essay responses and writing scores were used to develop the statistical models used for machine scoring the remaining online essay responses. All essays administered on paper tests were hand-scored. In addition, hand-scoring was required for a subset of math items administered on paper, generally equation items, where it was not possible to represent the item on paper in a way that allowed machine-scoring.

3. SUMMARY OF FALL 2015 OPERATIONAL TEST ADMINISTRATION

The following tests were administered in fall 2015:

- ELA (reading and writing) in grades 9 through 11
- Math in grades 9 through 11, following completion of Algebra I, Geometry, and Algebra II, or similar, coursework

Online administration of the AzMERIT occurred from October 26 through December 4, 2015. The paper version of the AzMERIT was administered between October 26 and November 6, 2015.

The scoring and reporting of the fall 2015 assessments used the items parameters calibrated following the spring 2015 administration and vertical scale and performance standards established in summer 2015. This section summarizes the operational test results for the fall 2015 administration of the AzMERIT.

3.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 3.1.1 and 3.1.2 by assessment and subgroup.¹⁷ Tables in Appendix D show the demographic composition of test takers by mode of test administration.

Exhibit 3.1.1 Number of Students Participating in ELA Assessments, by Test

Group	ELA 9	ELA 10	ELA 11
All Students	3798	4043	6093
Female	1765	1853	2965
Male	2033	2190	3128
Unknown	N/A	N/A	N/A
African American	237	235	404
Asian	68	85	126
Native Hawaiian/Pacific Islander	14	17	21
Hispanic/Latino	1711	1713	2886
American Indian or Alaskan	306	255	330
White	1376	1598	2198
Multiple Ethnicities	83	133	125
Limited English Proficiency	79	79	79
Special Education	523	523	523

¹⁷ Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio demographic and developmental characteristics.

Exhibit 3.1.2 Number of Students Participating in Math Assessments, by Test

Group	Algebra I	Geometry	Algebra II
All Students	7359	5887	5412
Female	3514	2730	2738
Male	3845	3157	2674
Unknown	N/A	N/A	N/A
African American	532	450	338
Asian	212	110	175
Native Hawaiian/Pacific Islander	35	19	26
Hispanic/Latino	3066	2811	2160
American Indian or Alaskan	285	222	245
White	3069	2098	2315
Multiple Ethnicities	155	176	153
Limited English Proficiency	118	91	75
Special Education	449	517	259

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibit 3.2.1.

Exhibit 3.2.1 Test Score Summary Statistics

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
ELA					
9	3798	2551.01	26.60	2454.00	2638.34
10	4043	2558.93	24.41	2458.47	2648.91
11	6093	2562.52	24.77	2474.47	2652.49
Math					
Algebra I	7359	3674.21	35.40	3577.00	3786.95
Geometry	5887	3671.83	28.24	3609.46	3802.43
Algebra II	5412	3692.78	33.45	3629.35	3828.53

The percentage of students in each performance level by grade and content area, as well as the percent of students at or above Proficient are presented in Exhibit 3.2.2.

Exhibit 3.2.2 Percentage of Students in Performance Levels

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
9	3798	56	25	17	2	19
10	4043	60	21	15	3	19
11	6093	60	19	16	5	20
Math						
Algebra I	7359	40	21	27	12	39
Geometry	5887	54	29	15	2	17
Algebra II	5412	46	24	24	6	30

3.3 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 3.3.1 and 3.3.2 present the number and percentage of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian Native Hispanic/Latino, American Indian, White, Multiple Ethnicities, limited English proficiency and special education.

Exhibit 3.3.1 Number of Students At Each Performance Level by Subgroups-Fall 2015

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
ELA													
9	Min Proficient	2127	883	1240	147	22	4	1163	214	537	28	301	81
	Part. Proficient	950	494	447	57	23	8	342	64	427	20	32	4
	Proficient	646	335	305	31	14	2	171	28	358	24	14	0
	Highly Proficient	76	53	41	0	9	0	17	0	55	11	0	0
10	Min Proficient	2426	1038	1402	155	29	6	1233	186	767	59	315	46
	Part. Proficient	849	408	438	42	23	4	308	43	384	40	33	2
	Proficient	606	315	307	31	28	4	154	20	352	29	11	1
	Highly Proficient	121	74	66	7	5	3	34	5	96	5	7	0
11	Min Proficient	3656	1720	1971	283	44	16	2049	241	967	66	455	76
	Part. Proficient	1158	623	563	65	38	1	462	50	528	30	42	2
	Proficient	975	474	469	48	33	4	289	30	528	24	26	1
	Highly Proficient	305	148	125	8	11	0	58	10	176	5	5	0

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
Mathematics													
Algebra I	Min Proficient	2944	1300	1653	261	32	9	1594	154	829	64	301	89
	Part. Proficient	1545	773	769	122	25	8	644	60	614	28	49	17
	Proficient	1987	1019	961	112	78	11	613	54	1074	40	67	13
	Highly Proficient	883	422	500	32	76	7	184	17	552	23	27	1
Geometry	Min Proficient	3179	1474	1673	294	36	7	1715	124	902	88	367	71
	Part. Proficient	1707	819	884	117	32	5	759	56	692	58	83	15
	Proficient	883	382	505	41	34	3	309	38	441	26	52	5
	Highly Proficient	118	55	95	0	8	4	28	4	84	4	16	0
Algebra II	Min Proficient	2490	1232	1257	166	39	10	1274	172	741	69	194	65
	Part. Proficient	1299	712	615	85	32	4	475	54	625	21	36	6
	Proficient	1299	657	642	81	74	6	346	17	741	47	26	4
	Highly Proficient	325	137	187	7	32	6	43	0	208	14	3	0

Exhibit 3.3.2 Percentage of Students At Each Performance Level by Subgroups—Fall 2015

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
ELA													
9	Min Proficient	56	50	61	62	32	29	68	70	39	34	86	95
	Part. Proficient	25	28	22	24	34	57	20	21	31	24	9	5
	Proficient	17	19	15	13	21	14	10	9	26	29	4	0
	Highly Proficient	2	3	2	0	13	0	1	0	4	13	0	0
10	Min Proficient	60	56	64	66	34	35	72	73	48	44	86	94
	Part. Proficient	21	22	20	18	27	24	18	17	24	30	9	4
	Proficient	15	17	14	13	33	24	9	8	22	22	3	2
	Highly Proficient	3	4	3	3	6	18	2	2	6	4	2	0
11	Min Proficient	60	58	63	70	35	76	71	73	44	53	87	96
	Part. Proficient	19	21	18	16	30	5	16	15	24	24	8	3
	Proficient	16	16	15	12	26	19	10	9	24	19	5	1
	Highly Proficient	5	5	4	2	9	0	2	3	8	4	1	0

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
Mathematics													
Algebra I	Min Proficient	40	37	43	49	15	26	52	54	27	41	67	75
	Part. Proficient	21	22	20	23	12	23	21	21	20	18	11	14
	Proficient	27	29	25	21	37	31	20	19	35	26	15	11
	Highly Proficient	12	12	13	6	36	20	6	6	18	15	6	1
Geometry	Min Proficient	54	54	53	66	33	37	61	56	43	50	71	78
	Part. Proficient	29	30	28	26	29	26	27	25	33	33	16	16
	Proficient	15	14	16	9	31	16	11	17	21	15	10	5
	Highly Proficient	2	2	3	0	7	21	1	2	4	2	3	0
Algebra II	Min Proficient	46	45	47	49	22	38	59	70	32	45	75	87
	Part. Proficient	24	26	23	25	18	15	22	22	27	14	14	8
	Proficient	24	24	24	24	42	23	16	7	32	31	10	5
	Highly Proficient	6	5	7	2	18	23	2	0	9	9	1	0

Note: Part. = Partially; Min. = Minimally; Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander.

3.4 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. The reliability evidence of the AzMERIT ELA and math are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance level classifications.¹⁸

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.¹⁹

¹⁸ Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

¹⁹ Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

3.4.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 3.4.1.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and math assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 3.4.1.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.87	707	0.91	1264
10/Geometry	0.86	589	0.81	735
11/Algebra II	0.86	599	0.87	1111

Note: Reliability ranges from 0 to 1. Variance is in scale score metric.

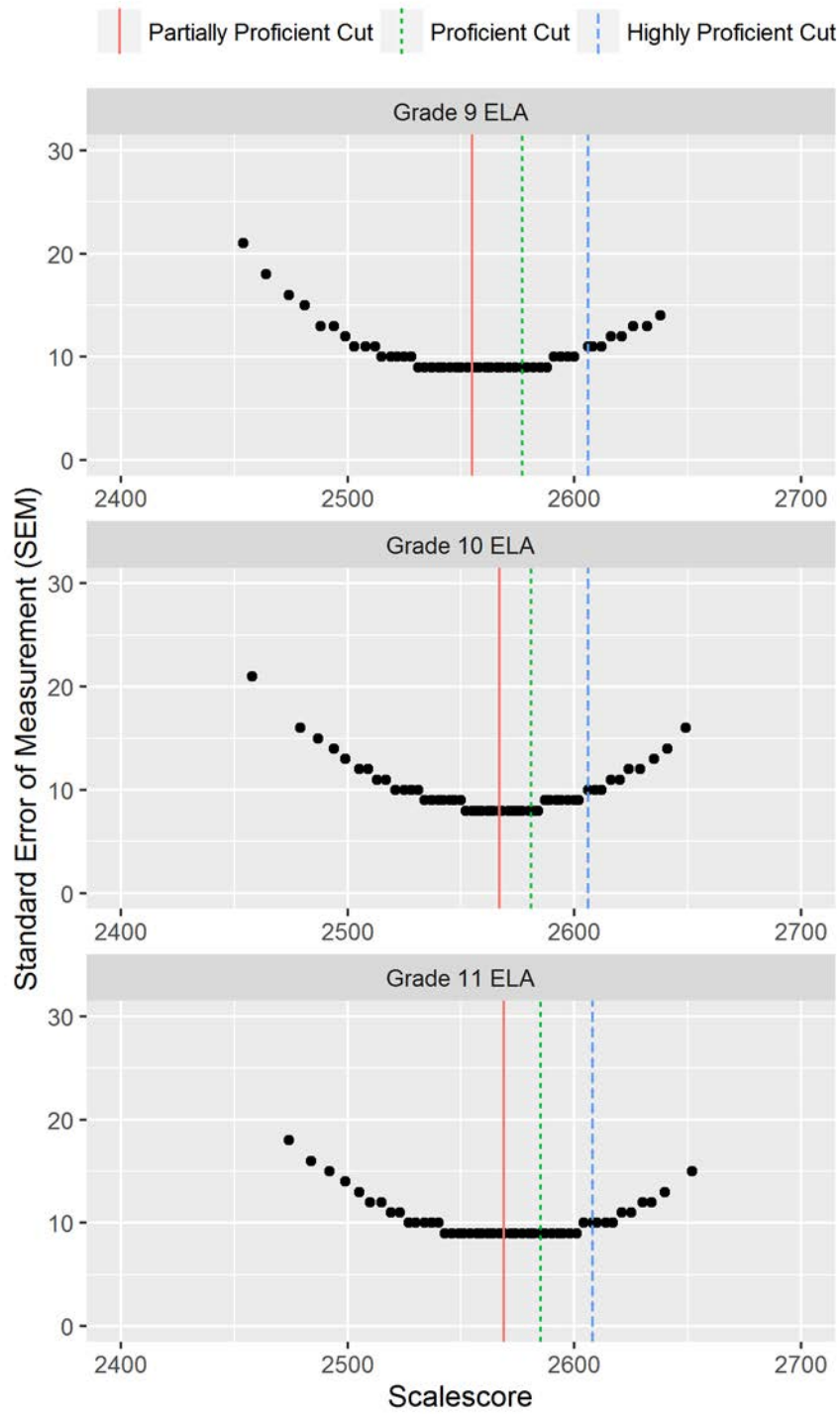
3.4.2 STANDARD ERROR OF MEASUREMENT

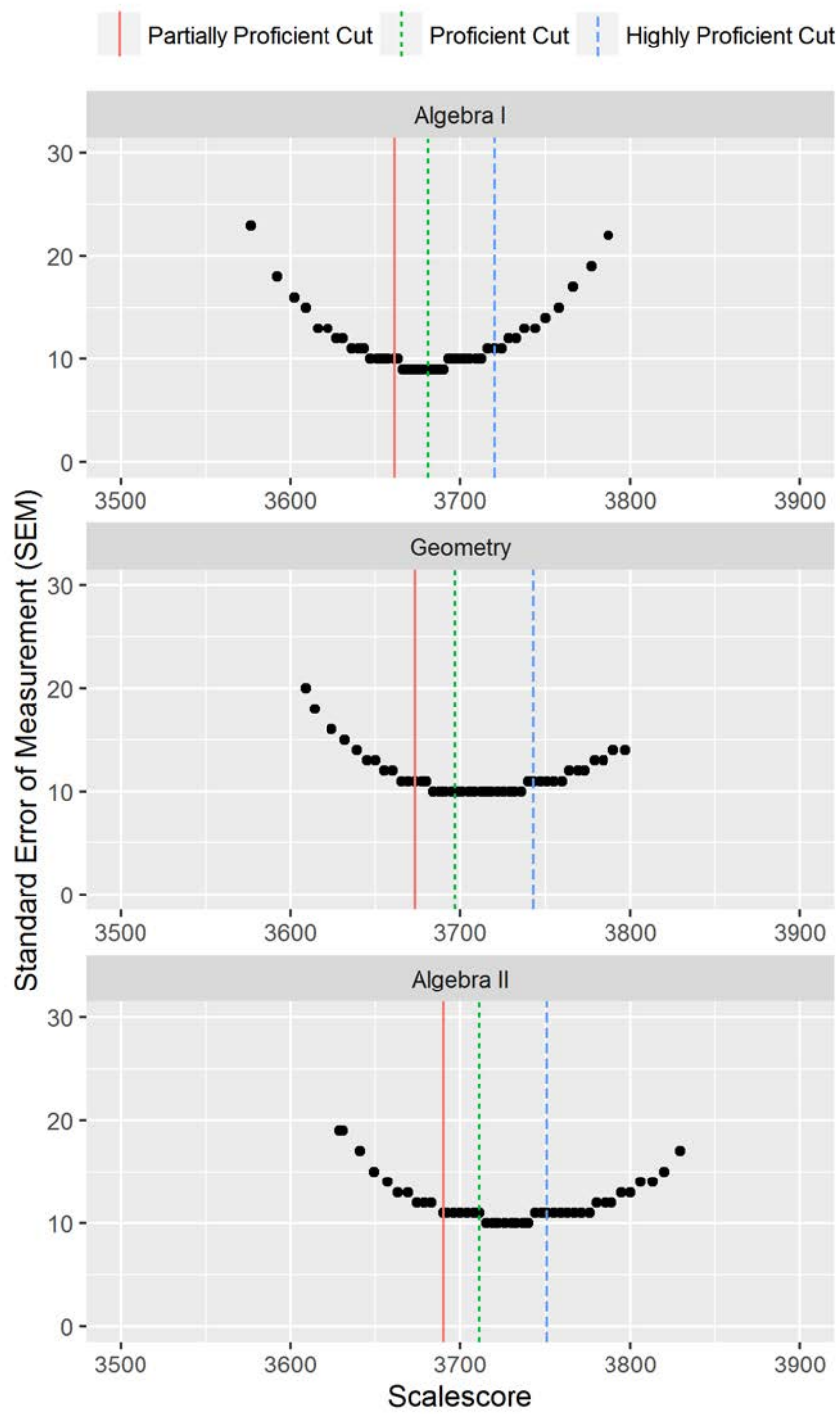
Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low and high performing students, the precision of test scores decreases near the tails of the ability distribution.

The figure in Exhibit 3.4.2 graphically presents the standard errors of measurement for the AzMERIT ELA and math assessments. Each figure also includes the location of the three AzMERIT performance standards. As the figures indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Partially Proficient and Proficient performance standards.²⁰ Test scores near the tails of the ability distribution are somewhat less precise, as expected. An SEM of .3 on the theta metric is consistent with an internal consistency of 0.9. The tables in Appendix F show the mean SEMs for students scoring in each of the performance levels on the AzMERIT reporting scale. While these tables indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance level classifications.

²⁰ Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 3.4.2.1 Overall Standard Error of Measurement for ELA and Math





3.4.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed in terms of the probabilities of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).²¹ This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications are typically estimated on a single-form test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for decision accuracy and decision consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Decision consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

3.4.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the above probabilities directly using the likelihood function. The likelihood function of the achievement attribute, designated θ , given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated theta is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 3.4.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications differ from their observed score classifications (e.g., a student whose true score results in a Proficient level classification, but is classified incorrectly as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

²¹ Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

Exhibit 3.4.4.1 Classification Accuracy

		Classification on a Form Actually Taken	
		At or Above the Cut Score	Below the Cut Score
Classification on True Score	At or Above the Cut Score	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

3.4.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 3.4.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 3.4.5.1 Classification Consistency

		Classification on the 2nd Form Taken	
		Above the Cut Score	Below the Cut Score
Classification on the 1st Form Taken	At or Above the Cut Score	N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

3.4.6 CLASSIFICATION RELIABILITY ESTIMATES

Exhibit 3.4.6.1 presents the classification accuracy and consistency indexes for the fall 2015 administration of AzMERIT. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Exhibit 3.4.6.1 Classification Accuracy and Consistency Indexes for Performance Standards

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
9	0.91	0.94	0.98	0.87	0.91	0.98
10	0.90	0.93	0.98	0.86	0.90	0.97
11	0.90	0.92	0.97	0.86	0.90	0.96
MATH						
Algebra I	0.91	0.93	0.96	0.88	0.90	0.94
Geometry	0.86	0.94	0.99	0.81	0.91	0.99
Algebra II	0.89	0.92	0.97	0.86	0.88	0.96

3.4.7 RELIABILITY FOR SUB-GROUPS IN THE POPULATION

Exhibit 3.4.7.1 and 3.4.7.2 shows the mean reliability for each of the subgroups: African Americans, Asian, Native Hawaiians/Pacific Islanders, Hispanic/Latinos, American Indian or Alaskans, Whites, Multiple Ethnicities and females and males (regardless of racial or ethnic group).²² Each racial and/or ethnic group was composed of approximately equal numbers of males and females. As the Exhibit indicates, internal consistency reliabilities are consistent across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups. Where reliability estimates are attenuated, there is an associated decrease in variance within the subgroup population, indicating that the decrease in reliability is likely due to a restriction in range.

Exhibit 3.4.7.1 Internal Consistency Reliability by Subgroup– ELA

	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.87	707	0.86	589	0.86	599
Female	0.87	674	0.86	544	0.85	571
Male	0.87	708	0.87	614	0.86	623
African American	0.85	614	0.86	573	0.85	601
Asian	0.89	884	0.84	480	0.85	554
Native Hawaiian/Pacific Islander	0.79	395	0.87	573	0.74	306
Hispanic/Latino	0.85	590	0.84	494	0.84	530
American Indian or Alaskan	0.83	524	0.85	536	0.82	489
White	0.87	651	0.87	593	0.85	576
Multiple Ethnicities	0.90	984	0.85	514	0.86	612
Limited English Proficiency	0.68	321	0.72	334	0.63	267
Special Education	0.78	457	0.77	383	0.80	479

Exhibit 3.4.7.2 Internal Consistency Reliability by Subgroup – Math

	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.91	1264	0.81	735	0.87	1111
Female	0.91	1204	0.79	643	0.86	997
Male	0.91	1318	0.83	811	0.88	1227
African American	0.89	997	0.75	583	0.84	910
Asian	0.91	1557	0.87	1021	0.89	1206
Native Hawaiian/Pacific Islander	0.90	1151	0.94	2003	0.91	1549
Hispanic/Latino	0.89	999	0.78	636	0.82	898
American Indian or Alaskan	0.88	948	0.83	814	0.71	605
White	0.91	1267	0.84	807	0.88	1090
Multiple Ethnicities	0.91	1180	0.78	628	0.89	1273
Limited English Proficiency	0.83	656	0.67	461	0.66	553
Special Education	0.86	861	0.65	482	0.75	764

²² Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

3.4.8 SUBSCALE RELIABILITY

Coefficient alpha internal consistency reliability estimates associated with the subscales for the Fall 2015 operational forms are presented in Exhibit 3.4.8.1-3.4.8.3. As indicated in the Exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT. The only exception is the Modeling with Geometry strand in Geometry test. The reason is that nearly 40 percent of students received raw score 0 in this strand. Using marginal reliability is not a good choice because the standard deviation of the observed score is too arbitrary given the rule of estimating the extreme scores.

Exhibit 3.4.8.1 Subscale Reliabilities – ELA Grades 9-11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.76	0.66	0.73
Grade 10	0.69	0.70	0.71
Grade 11	0.70	0.59	0.74

Exhibit 3.4.8.2 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.83	0.78	0.48
Algebra II	0.73	0.60	0.66

Exhibit 3.4.8.3 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.43	0.54	-0.20	0.60

3.5 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 3.5.1-3.5.3. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.²³ The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

²³ Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.

Where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y .

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Exhibit 3.5.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 9 to 11

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.63		0.89	
	Writing & Language	0.63	0.55	0.84	0.80
10	Literature	0.65		0.90	
	Writing & Language	0.56	0.57	0.77	0.81
11	Literature	0.61		0.91	
	Writing & Language	0.63	0.58	0.83	0.87

Exhibit 3.5.2 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.80		0.99	
	Statistics	0.61	0.61	0.97	0.99
Algebra II	Functions	0.66		1.00	
	Statistics	0.71	0.65	1.00	1.00

Exhibit 3.5.3 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.49			1.00		
	Modeling with Geometry (MG)	0.47	0.49		1.00	1.00	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.49	0.58	0.54	0.97	1.00	1.00

4. SUMMARY OF SPRING 2016 OPERATIONAL TEST ADMINISTRATION

The following AzMERIT assessments were administered in spring 2016:

- ELA (reading and writing) in grades 3 through 11
- Math in grades 3 through 8, Algebra I, Geometry, and Algebra II.

Online administration of the AzMERIT occurred from March 28 through May 6, 2016. The paper version of the AzMERIT was administered between March 28 and April 8, 2016.

The Math tests were scored using pre-equated item parameter estimates derived from the initial online administration of the AzMERIT tests in spring 2015. Thus, no post-equating activities were conducted prior to the scoring and reporting of the Math tests. In spring 2016, in each ELA online assessment, students were randomly assigned one of six writing prompts for administration. Following the test administration, all operational items including reading and writing items were concurrently calibrated, and then linked back to the AzMERIT bank scale using the mean-mean equating method. The post-equated item parameters were used for the final scoring and reporting. This section summarizes the operational test results for the spring 2016 administration of the AzMERIT. Detailed descriptions of procedures for item and test development, test administration, scaling, equating, and scoring are presented in subsequent sections.

4.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 4.1.1 and 4.1.2 by assessment and subgroup.²⁴ We note that some students participated in an end-of-course assessment rather than a grade level assessment, especially in grade 8 where more advanced students are enrolled in Algebra I courses. The tables in Appendix G show the demographic composition of test takers by mode of test administration.

Exhibit 4.1.1 Number of Students Participating in ELA Assessments, by Test Name

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
All Students	87793	86325	85425	84651	84138	82779	80130	73403	64834
Female	43134	42452	42102	41190	41418	40757	39488	36507	32208
Male	44645	43861	43315	43456	42704	42011	40594	36851	32596
Unknown	14	12	8	5	16	11	48	45	30
African American	4437	4432	4430	4518	4487	4555	4305	3995	3331
Asian	2422	2322	2466	2542	2419	2409	2423	2302	2120
Native Hawaiian/Pacific Islander	313	349	277	272	245	228	288	207	217
Hispanic/Latino	40327	39205	38172	37569	37123	36622	35381	31489	27202
American Indian or Alaskan	4354	4333	4426	4293	4001	4063	3900	3574	3153
White	33223	33150	33418	33434	33946	33145	32485	30583	27744
Multiple Ethnicities	2703	2522	2228	2018	1901	1745	1297	1198	1031
Limited English Proficiency	8491	8311	6470	4916	4105	3347	3487	2616	1755
Special Education	9581	10183	10204	9549	8903	8553	6946	6024	4996
Free/Reduced Lunch	37747	36855	36060	35466	34799	33432	30003	26487	22191

²⁴ Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio demographic and developmental characteristics.

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
Accommodation	13439	13017	10931	9227	7612	6470	3111	2191	1649

Exhibit 4.1.2 Number of Students Participating in Math Assessments, by Test Name

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
All Students	88303	86711	85719	84675	81829	69858	82623	71654	60900
Female	43349	42607	42221	41173	40307	34103	40425	35652	30768
Male	44941	44092	43490	43496	41506	35736	42130	35949	30090
Unknown Gender	13	12	8	6	16	19	68	53	42
African American	4488	4491	4454	4543	4437	4204	4506	3662	3196
Asian	2435	2341	2460	2353	1884	1415	2402	2356	1951
Native Hawaiian/Pacific Islander	316	348	279	271	243	200	261	222	192
Hispanic/Latino	40550	39355	38298	37738	36940	33590	37028	30642	25612
American Indian or Alaskan	4404	4353	4451	4333	4074	3803	4453	3595	2920
White	33371	33268	33528	33398	32395	25159	32569	29962	26009
Multiple Ethnicities	2726	2543	2241	2033	1840	1467	1333	1159	977
Limited English Proficiency	8622	8402	6532	4967	4143	3153	3747	2857	1662
Special Education	9731	10265	10299	9628	8993	8456	7416	5319	3576
Free/Reduced Lunch	37157	36790	35932	35315	34687	30530	31860	26108	21308
Accommodation	13518	13134	11006	9309	7611	6402	3068	1886	1247

4.2 CLASSICAL ITEM ANALYSIS

Because AzMERIT is an online assessment system, classical item analysis statistics for multiple-choice (MC) and constructed-response (CR) items reported here are calculated based on all online student responses. Classical item analysis statistics are used to monitor item behavior and investigate irregularities in item scoring throughout the test window for online assessments, and following processing of answer documents for paper test administrations. Classical item analyses ensure that the items function as intended with respect to the underlying scales. For online and paper test administrations, quality assurance reports provide the required item and test statistics for each multiple-choice and constructed-response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item during test administration. Key statistics computed and examined include biserial/polyserial correlations for item discrimination, biserial correlations for distractors for selected response items, and proportion correct for item difficulty.

The biserial/polyserial correlations indicate the extent to which each item differentiated between those examinees who possessed the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The biserial correlation for multiple-choice items is calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, the mean total number correct for student scoring within each of the possible score categories is used. Items are flagged for review by test development experts if the biserial correlation for the keyed (correct) response is less than .25, or changed from previous administration. For multiple-choice items, we also compute the biserial correlation for each of the distractor response options.

The proportion correct score is the average number of available points achieved by students on the item. For dichotomous items, this is simply the proportion of students responding correctly. For polytomous items, the average score on the item is divided by the points available to produce a comparable index. The proportion correct score is commonly referred to as the p -value.

Exhibit 4.2.1 presents the average proportion of students responding correctly and average point biserial/polyserial correlations from the spring 2016 online administration of AzMERIT. As indicated in Exhibit

4.2.1, ELA items were are somewhat harder than the math items for students in Grades 3 through 6, where this trend is reversed in Grades 7 and above, with items on the ELA assessments, on average, being easier than items on the math assessments. While mean difficulty of ELA items is relatively consistent across grade level assessments, average difficulty of math items increases across grade level and course assessments. The proportion of students responding correctly to test items in the end of course assessments in math was relatively quite low. Mean biserial correlations for the grade level and end of course assessments are reasonably high and consistent across assessments. Exhibit 4.2.2 shows the number of items flagged for proportion correct value, biserial/polyserial correlation, distractor biserial/polyserial, and DIF categories for the operational items in the spring 2016 online forms. The flagging criteria are presented in section 5.4.1 and 5.4.3.

Exhibit 4.2.1 Average Proportion Correct and Point Biserial Correlations for Operational Test Items Administered Online

Grade	Average <i>p</i> -Value	<i>p</i> -Value SD	Average Point- Biserial	Point- Biserial SD
ELA				
3	0.48	0.18	0.48	0.13
4	0.44	0.23	0.47	0.11
5	0.48	0.21	0.49	0.13
6	0.46	0.21	0.50	0.13
7	0.47	0.18	0.48	0.12
8	0.50	0.18	0.49	0.13
9	0.49	0.18	0.46	0.13
10	0.48	0.18	0.44	0.15
11	0.48	0.15	0.48	0.14
Math				
3	0.62	0.20	0.50	0.10
4	0.56	0.19	0.48	0.10
5	0.51	0.20	0.48	0.11
6	0.49	0.21	0.49	0.10
7	0.46	0.19	0.45	0.12
8	0.40	0.21	0.43	0.12
Algebra I	0.41	0.16	0.45	0.12
Geometry	0.33	0.17	0.44	0.09
Algebra II	0.28	0.18	0.41	0.09

Exhibit 4.2.2 Number of Items Flagged For P-value, Biserial/Polyserial or DIF for Operational Test Items Administered Online

Grade	Number of Flagged Operational Items			
	Proportion Correct	Biserial/Polyserial Correlation	Biserial Correlation for Distractor	Differential Item Functioning
ELA				
3	0	0	0	7
4	0	0	0	9
5	0	0	0	1
6	0	0	1	7
7	1	0	1	9
8	0	0	0	2
9	0	0	0	3
10	0	1	2	4
11	0	0	0	4
MATH				
3	1	0	0	1
4	0	0	0	1
5	0	0	0	0
6	0	0	1	0
7	2	1	0	0
8	0	1	0	0
Algl	2	1	2	0
Geo	3	0	2	0
AlgII	1	1	1	2

4.3 ITEM RESPONSE THEORY ANALYSIS

Calibration is the process by which the statistical relationship between item responses and the underlying measurement construct is estimated. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^n P(z_j|\theta),$$

where Z represents the vector of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function $P(Z)$. The one-parameter model (also known as the Rasch model) is used to calibrate dichotomously scored AzMERIT items and takes the form

$$P(x_j = 1|\theta_k, b_j) = \frac{1}{1+e^{(\theta_k-b_j)}} = P_{j1}(\theta_k).$$

The b parameter is often called the *location* or *difficulty* parameter—the greater the value of b , the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

$$P(x_j = i|\theta_k, b_{j0} \dots b_{jm_j-1}) = \frac{e^{\sum_{v=0}^i (\theta_k - b_{jv})}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^g (\theta_k - b_{jv})}}.$$

The tables in Appendix H provide Rasch and Masters' partial credit model item parameter estimates for the spring 2016 operational test items. Since AzMERIT is an online assessment system, bank item parameters were estimated based only on online responses to test items. Exhibit 4.3.1 presents the mean and standard deviation of the Rasch item parameters by item type for each test for items administered online. Item types include traditional four-option multiple choice (MC) items, technology-enhanced (TE) selected response items which may require students to select one or more options, and machine-scored constructed response (MSCR) items for which students' constructed responses are scored electronically using explicit rubrics. In addition, the average Rasch difficulty is presented for each scoring dimension of the writing prompt administered at each grade. As illustrated in Exhibit 4.3.1, selected-response items are, on average, less difficult than the constructed-response item types. Within the constructed response items, Evidence and Elaboration within the writing prompts was on average, consistently found to be the most difficult.

Exhibit 4.3.1 Rasch Summary Statistics by Item Type for Items Administered Online

Grade/Course	MC			TE Selected Response			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
ELA												
3	28	-0.01	0.70	10	-0.12	1.26	3	0.81	0.93	1.61	1.67	-1.18
4	23	-0.14	0.75	13	0.60	1.26	5	1.01	1.12	3.70	4.09	0.00
5	19	-0.06	0.67	13	0.19	1.24	9	1.40	1.36	2.38	2.90	-0.94
6	26	-0.18	0.80	9	0.45	0.78	6	0.96	1.12	2.29	2.89	-1.27
7	22	0.12	0.69	16	0.48	0.68	3	0.80	0.72	2.36	2.66	-1.42
8	25	-0.18	0.75	10	0.21	0.97	6	1.09	1.03	1.03	1.21	-1.62
9	24	-0.13	0.68	12	0.21	0.78	7	1.01	0.36	1.35	1.63	-1.67
10	30	-0.06	0.58	10	0.57	1.27	3	1.44	0.43	0.81	1.18	-2.09
11	23	-0.09	0.38	15	0.13	0.94	5	0.17	0.55	0.25	0.73	-1.90

Grade/Course	MC			TE Selected Response			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
Mathematics												
3	16	-0.52	1.39	1	2.28	NA	28	0.23	1.33	--	--	--
4	13	-0.58	1.10	7	0.29	1.22	25	0.29	1.24	--	--	--
5	13	-0.47	0.79	4	-0.50	1.22	28	0.44	1.35	--	--	--
6	12	-0.51	1.09	5	1.16	1.48	30	0.00	1.41	--	--	--
7	24	-0.35	1.03	3	0.50	0.90	20	0.65	0.92	--	--	--
8	18	-0.91	0.93	3	0.25	0.74	26	0.51	1.35	--	--	--
Algebra I	28	-0.14	0.78	2	0.96	1.22	17	0.26	1.00	--	--	--
Geometry	19	-0.70	0.89	5	0.69	0.91	23	0.43	1.27	--	--	--
Algebra II	20	-1.04	0.77	4	0.77	0.70	23	0.78	0.92	--	--	--

Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.7 or greater than 1.3. Exhibit 4.3.2 summarizes the number of online administered operational test items with Infit and Outfit statistics within the range of .7 to 1.3.

Exhibit 4.3.2 Summary of Item Fit Statistics for Items Administered Online

Grade/ Course	Infit			Outfit		
	Below	Between	Above	Below	Between	Above
	0.7	.7-1.3	1.3	0.7	.7-1.3	1.3
ELA						
3	0	59	0	0	57	2
4	0	59	0	0	55	4
5	0	58	1	1	56	2
6	0	59	0	1	56	2
7	0	59	0	1	58	0
8	0	59	0	6	51	2
9	0	61	0	5	55	1
10	0	61	0	5	56	0
11	0	61	0	4	57	0

Grade/ Course	Infit			Outfit		
	Below	Between	Above	Below	Between	Above
	0.7	.7-1.3	1.3	0.7	.7-1.3	1.3
Mathematics						
3	0	44	1	0	36	9
4	0	42	3	3	35	7
5	0	42	3	1	34	10
6	0	46	1	3	35	9
7	0	45	2	4	36	7
8	0	46	1	3	36	8
Algebra I	0	45	2	3	37	7
Geometry	0	47	0	3	39	5
Algebra II	0	47	0	5	40	2

4.4 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibit 4.4.1. The AzMERIT bank scale was established based on the spring 2015 assessments in which the item calibrations were centered on items rather than persons, resulting in operational test forms with mean difficulty of zero and standard deviation of one. Because calibrations were not centered on persons, the standard deviation of ability estimates is not expected to be 30, as might be implied by the scaling transformation.

Exhibit 4.4.1 Test Score Summary Statistics

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
ELA					
3	87793	2501.38	31.81	2605	2395
4	86325	2517.62	33.89	2610	2400
5	85425	2536.70	34.41	2629	2419
6	84651	2540.77	33.74	2641	2431
7	84138	2551.77	31.12	2648	2438
8	82779	2555.26	32.38	2658	2448
9	80130	2564.11	31.43	2664	2454
10	73403	2564.75	29.05	2668	2458
11	64834	2567.09	31.49	2675	2465
Math					
3	88303	3524.19	44.86	3605	3395
4	86711	3552.27	40.62	3645	3435
5	85719	3587.75	41.57	3688	3478
6	84675	3615.26	42.60	3722	3512
7	81829	3632.28	35.61	3739	3529
8	69858	3651.11	36.11	3776	3566
Algebra I	82623	3669.80	34.18	3787	3577

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
Geometry	71654	3684.30	36.07	3819	3609
Algebra II	60900	3695.35	33.00	3839	3629

The percentage of students in each performance level by grade and content area, as well as the percent of students at or above Proficient are presented in Exhibit 4.4.2.

Exhibit 4.4.2 Percentage of Students in Performance Levels

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	87793	45	14	29	12	41
4	86325	40	14	34	12	46
5	85425	32	23	32	13	45
6	84651	41	21	31	6	38
7	84138	38	21	35	6	41
8	82779	44	23	25	7	33
9	80130	37	28	25	9	35
10	73403	54	17	20	10	30
11	64834	53	17	19	11	30
Math						
3	88303	25	30	28	17	45
4	86711	29	26	34	10	44
5	85719	27	27	30	15	46
6	84675	37	24	24	14	39
7	81829	46	23	22	9	31
8	69858	51	23	17	9	26
Algebra I	82623	45	19	27	9	36
Geometry	71654	41	24	28	8	35
Algebra II	60900	46	25	22	7	29

4.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 4.5.1 and 4.5.2 present the number and percentage, respectively, of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian Native, Hispanic/Latino, American Indian, White, and Multiple Ethnicities, other demographic information such as special education (SPED) and limited English proficiency (LEP).

Exhibit 4.5.1 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
ELA															
3	Minimally Proficient	39507	17685	21430	2485	509	141	22583	2874	9635	973	7377	7217	21138	11128
	Partially Proficient	12291	6039	6250	666	315	44	5646	610	4651	378	766	594	5662	1096
	Proficient	25460	13372	12054	1065	896	100	9275	784	12293	892	1054	509	8682	1098
	Highly Proficient	10535	6039	4911	266	678	31	2420	131	6645	460	383	85	2265	117
4	Minimally Proficient	34530	1596	1022	185	7449	1647	16907	1614	10613	13597	7841	7064	18796	10717
	Partially Proficient	12086	620	325	49	3921	693	4973	353	5519	5702	815	665	5528	1091
	Proficient	29351	1640	743	94	17250	1517	9282	479	18254	17544	1324	582	10688	1125
	Highly Proficient	10359	620	232	21	10585	477	1989	76	8490	7018	305	0	2211	84
5	Minimally Proficient	27336	1152	912	119	4581	1151	13367	1225	7999	10829	7551	5111	14785	8603
	Partially Proficient	19648	1019	592	66	6108	1151	8689	557	8420	9529	1429	971	9376	1633
	Proficient	27336	1551	715	75	14887	1416	9023	401	16841	15593	1020	324	9736	635
	Highly Proficient	11105	709	271	17	12597	708	2339	67	8841	7797	306	65	2524	60
6	Minimally Proficient	34707	1672	1144	144	6762	1674	17720	1352	10298	14340	7735	4424	18797	8097
	Partially Proficient	17777	994	534	60	6387	1116	7355	383	8650	9126	955	344	7803	768
	Proficient	26242	1536	737	60	17657	1202	7355	262	17712	16513	668	147	7803	338
	Highly Proficient	5079	361	127	8	7138	301	669	20	4119	3476	95	0	709	24
7	Minimally Proficient	31972	1481	1040	115	6682	1800	16294	1236	9940	13238	7211	3653	17052	6667
	Partially Proficient	17669	987	508	54	5940	960	7468	380	8284	8541	890	287	7656	616
	Proficient	29448	1705	774	69	17819	1080	9165	285	19052	17509	712	164	9048	313
	Highly Proficient	5048	359	121	7	6311	160	1018	19	4142	3416	89	0	696	16

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
8	Minimally Proficient	36423	1776	1180	125	7324	1828	17898	1239	12227	15964	7270	2878	18722	5884
	Partially Proficient	19039	1093	530	50	6958	1016	7623	314	9782	10083	770	234	7689	395
	Proficient	20695	1275	554	46	13550	975	6298	175	13857	11763	428	201	6018	174
	Highly Proficient	5795	410	145	9	8789	244	1326	17	4891	3781	86	33	1003	17
9	Minimally Proficient	29648	1378	1042	144	6015	1209	15268	791	9872	12178	5487	2650	15002	2716
	Partially Proficient	22436	1248	654	78	7784	1287	9421	337	10662	10554	972	523	8401	283
	Proficient	20033	1205	557	55	12383	1053	6497	143	13426	12178	417	279	5401	96
	Highly Proficient	7212	474	194	14	9553	390	1624	26	5923	5277	69	35	1200	16
10	Minimally Proficient	39638	1998	1335	137	9132	2073	19879	934	14603	16951	5241	2171	18011	2030
	Partially Proficient	12479	719	368	31	4408	429	4893	144	6936	6265	422	235	3973	83
	Proficient	14681	879	414	29	8817	715	4282	96	9857	8476	301	183	3443	60
	Highly Proficient	7340	439	184	10	9132	357	1529	24	5476	5159	60	52	1059	18
11	Minimally Proficient	34362	1566	1230	143	7889	1860	17756	784	12883	14342	4396	1404	14868	1496
	Partially Proficient	11022	600	339	33	4080	441	4439	144	6120	5215	300	176	3329	85
	Proficient	12318	733	360	33	6801	568	4162	93	8052	8149	200	140	2885	51
	Highly Proficient	7132	400	191	11	8161	284	1387	21	5153	4889	50	35	1110	17
Mathematics															
3	Minimally Proficient	22076	10404	11685	1705	195	82	12976	1762	5006	572	5449	4828	12262	7453
	Partially Proficient	26491	13438	12583	1391	390	95	13787	1585	8343	709	2335	2587	12633	3947
	Proficient	24725	12571	12583	987	779	95	10138	837	11346	872	1362	1035	9661	1782
	Highly Proficient	15012	6936	8089	404	1071	47	4055	220	8676	572	584	172	2973	336
4	Minimally Proficient	25146	12356	13228	1976	211	94	14955	2046	5656	585	6570	5461	14348	8586
	Partially Proficient	22545	11504	11464	1257	351	90	11413	1306	7652	661	1950	2016	10669	2995
	Proficient	29482	14486	14550	1078	1053	129	11019	914	14305	966	1437	840	10301	1436
	Highly Proficient	8671	4261	4850	180	726	31	1968	131	5656	331	308	84	1840	117
5	Minimally Proficient	23144	10555	13047	1782	172	67	13404	2092	5364	493	6797	4311	12936	7504
	Partially Proficient	23144	11822	11307	1381	369	70	11489	1291	8047	583	2060	1502	10780	2425

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
	Proficient	25716	13511	12612	980	886	103	9957	890	12405	740	1133	588	9342	924
	Highly Proficient	12858	6333	6524	312	1009	39	3447	178	7376	426	309	131	2875	153
6	Minimally Proficient	31330	14411	16963	2408	353	87	17737	2556	7682	630	7510	3974	17304	7389
	Partially Proficient	20322	10293	10004	1045	376	81	9435	997	7682	488	1252	695	8829	1285
	Proficient	20322	10705	10004	772	753	60	7548	607	10353	549	674	248	6710	514
	Highly Proficient	11855	5764	6524	273	871	46	3019	173	7682	366	289	50	2472	121
7	Minimally Proficient	37641	17735	19923	2751	433	114	20686	2811	10366	736	7644	3687	19772	6814
	Partially Proficient	18821	9674	8716	887	377	68	8127	774	7775	460	809	331	7631	502
	Proficient	18002	9271	8716	621	565	44	6280	407	9395	442	450	124	5550	232
	Highly Proficient	7365	4031	3736	177	509	17	1847	81	4859	202	90	41	1387	63
8	Minimally Proficient	35628	16710	18940	2649	354	102	19482	2700	9560	719	7272	2743	18013	5694
	Partially Proficient	16067	8185	7862	883	311	42	7054	685	6541	337	761	252	6717	445
	Proficient	11876	6139	5718	547	354	32	4703	304	5787	264	338	126	3969	196
	Highly Proficient	6287	3069	3216	168	382	26	2015	114	3271	147	169	32	1832	67
Algebra I	Minimally Proficient	37180	16574	19801	2613	408	107	19995	3028	10096	507	6155	2848	17842	2588
	Partially Proficient	15698	8489	7583	856	336	73	7406	757	6514	280	742	412	6053	272
	Proficient	22308	11723	10533	901	961	65	8146	623	11399	400	519	412	6372	180
	Highly Proficient	7436	3638	3792	135	697	16	1851	89	4560	147	74	75	1593	28
Geometry	Minimally Proficient	29378	13904	15458	1977	377	100	15627	2193	8389	406	4149	1914	13576	1564
	Partially Proficient	17197	8913	8268	916	424	44	7661	863	6891	290	745	514	6266	230
	Proficient	20063	10339	9347	659	1013	62	6128	503	10786	336	372	371	5222	81
	Highly Proficient	5732	2496	2876	110	518	13	919	72	3595	127	53	57	783	11
Algebra II	Minimally Proficient	28014	13846	14443	1918	429	104	14343	1956	8843	410	2897	1014	12998	1055
	Partially Proficient	15225	8000	6921	735	390	46	6403	672	6502	234	429	382	4901	136
	Proficient	13398	7077	6319	479	683	35	4098	263	7803	234	179	216	2983	50
	Highly Proficient	4263	1846	2407	96	449	8	768	29	2861	98	36	50	426	6

Exhibit 4.5.2 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information.

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
ELA															
3	Minimally Proficient	45	41	48	56	21	45	56	66	29	36	77	85	56	83
	Partially Proficient	14	14	14	15	13	14	14	14	14	14	8	7	15	8
	Proficient	29	31	27	24	37	32	23	18	37	33	11	6	23	8
	Highly Proficient	12	14	11	6	28	10	6	3	20	17	4	1	6	1
4	Minimally Proficient	40	36	44	53	19	38	51	64	25	31	77	85	51	82
	Partially Proficient	14	14	14	14	10	16	15	14	13	13	8	8	15	8
	Proficient	34	37	32	27	44	35	28	19	43	40	13	7	29	9
	Highly Proficient	12	14	10	6	27	11	6	3	20	16	3	0	6	1
5	Minimally Proficient	32	26	37	43	12	26	40	55	19	25	74	79	41	79
	Partially Proficient	23	23	24	24	16	26	26	25	20	22	14	15	26	15
	Proficient	32	35	29	27	39	32	27	18	40	36	10	5	27	6
	Highly Proficient	13	16	11	6	33	16	7	3	21	18	3	1	7	1
6	Minimally Proficient	41	37	45	53	18	39	53	67	25	33	81	90	53	88
	Partially Proficient	21	22	21	22	17	26	22	19	21	21	10	7	22	8
	Proficient	31	34	29	22	47	28	22	13	43	38	7	3	22	4
	Highly Proficient	6	8	5	3	19	7	2	1	10	8	1	0	2	0
7	Minimally Proficient	38	33	43	47	18	45	48	65	24	31	81	89	49	88
	Partially Proficient	21	22	21	22	16	24	22	20	20	20	10	7	22	8
	Proficient	35	38	32	28	48	27	27	15	46	41	8	4	26	4
	Highly Proficient	6	8	5	3	17	4	3	1	10	8	1	0	2	0
8	Minimally Proficient	44	39	49	55	20	45	54	71	30	38	85	86	56	91
	Partially Proficient	23	24	22	22	19	25	23	18	24	24	9	7	23	6
	Proficient	25	28	23	20	37	24	19	10	34	28	5	6	18	3
	Highly Proficient	7	9	6	4	24	6	4	1	12	9	1	1	3	0
9	Minimally Proficient	37	32	43	50	17	31	47	61	25	30	79	76	50	87
	Partially Proficient	28	29	27	27	22	33	29	26	27	26	14	15	28	9
	Proficient	25	28	23	19	35	27	20	11	34	30	6	8	18	3
	Highly Proficient	9	11	8	5	27	10	5	2	15	13	1	1	4	1
10	Minimally Proficient	54	50	58	66	29	58	65	78	40	46	87	83	68	93
	Partially Proficient	17	18	16	15	14	12	16	12	19	17	7	9	15	4
	Proficient	20	22	18	14	28	20	14	8	27	23	5	7	13	3

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	Highly Proficient	10	11	8	5	29	10	5	2	15	14	1	2	4	1
11	Minimally Proficient	53	47	58	66	29	59	64	76	40	44	88	80	67	91
	Partially Proficient	17	18	16	15	15	14	16	14	19	16	6	10	15	5
	Proficient	19	22	17	15	25	18	15	9	25	25	4	8	13	3
	Highly Proficient	11	12	9	5	30	9	5	2	16	15	1	2	5	1
Math															
3	Minimally Proficient	25	24	26	38	8	26	32	40	15	21	56	56	33	55
	Partially Proficient	30	31	28	31	16	30	34	36	25	26	24	30	34	29
	Proficient	28	29	28	22	32	30	25	19	34	32	14	12	26	13
	Highly Proficient	17	16	18	9	44	15	10	5	26	21	6	2	8	2
4	Minimally Proficient	29	29	30	44	9	27	38	47	17	23	64	65	39	65
	Partially Proficient	26	27	26	28	15	26	29	30	23	26	19	24	29	23
	Proficient	34	34	33	24	45	37	28	21	43	38	14	10	28	11
	Highly Proficient	10	10	11	4	31	9	5	3	17	13	3	1	5	1
5	Minimally Proficient	27	25	30	40	7	24	35	47	16	22	66	66	36	68
	Partially Proficient	27	28	26	31	15	25	30	29	24	26	20	23	30	22
	Proficient	30	32	29	22	36	37	26	20	37	33	11	9	26	8
	Highly Proficient	15	15	15	7	41	14	9	4	22	19	3	2	8	1
6	Minimally Proficient	37	35	39	53	15	32	47	59	23	31	78	80	49	79
	Partially Proficient	24	25	23	23	16	30	25	23	23	24	13	14	25	14
	Proficient	24	26	23	17	32	22	20	14	31	27	7	5	19	6
	Highly Proficient	14	14	15	6	37	17	8	4	23	18	3	1	7	1
7	Minimally Proficient	46	44	48	62	23	47	56	69	32	40	85	89	57	90
	Partially Proficient	23	24	21	20	20	28	22	19	24	25	9	8	22	7
	Proficient	22	23	21	14	30	18	17	10	29	24	5	3	16	3
	Highly Proficient	9	10	9	4	27	7	5	2	15	11	1	1	4	1
8	Minimally Proficient	51	49	53	63	25	51	58	71	38	49	86	87	59	89
	Partially Proficient	23	24	22	21	22	21	21	18	26	23	9	8	22	7
	Proficient	17	18	16	13	25	16	14	8	23	18	4	4	13	3
	Highly Proficient	9	9	9	4	27	13	6	3	13	10	2	1	6	1
Algebra I	Minimally Proficient	45	41	47	58	17	41	54	68	31	38	83	76	56	84
	Partially Proficient	19	21	18	19	14	28	20	17	20	21	10	11	19	9
	Proficient	27	29	25	20	40	25	22	14	35	30	4	11	20	6

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
Geometry	Highly Proficient	9	9	9	3	29	6	5	2	14	11	1	2	5	1
	Minimally Proficient	41	39	43	54	16	45	51	61	28	35	78	67	52	83
	Partially Proficient	24	25	23	25	18	20	25	24	23	25	14	18	24	12
	Proficient	28	29	26	18	43	28	20	14	36	29	7	13	20	4
Algebra II	Highly Proficient	8	7	8	3	22	6	3	2	12	11	1	2	3	1
	Minimally Proficient	46	45	48	60	22	54	56	67	34	42	81	61	61	85
	Partially Proficient	25	26	23	23	20	24	25	23	25	24	12	23	23	11
	Proficient	22	23	21	15	35	18	16	9	30	24	5	13	14	4
	Highly Proficient	7	6	8	3	23	4	3	1	11	10	1	3	2	0

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch.

4.6 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. Evidence of the reliability of AzMERIT ELA and math scores are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance level classifications.²⁵

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability, such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.²⁶

²⁵ Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

²⁶ Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

4.6.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 4.6.1.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and math assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 4.6.1.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade	ELA		MATH	
	Reliability	Variance	Reliability	Variance
G3	0.89	999	0.91	1711
G4	0.89	1099	0.92	1518
G5	0.90	1154	0.92	1620
G6	0.90	1089	0.92	1685
G7	0.89	973	0.91	1263
G8	0.90	1016	0.90	1234
G9E / Algebra I	0.89	909	0.90	1126
G10E / Geometry	0.87	764	0.88	1216
G11E / Algebra II	0.89	927	0.85	1004

Note: Reliability ranges from 0 to 1. The variance is in scale score metric.

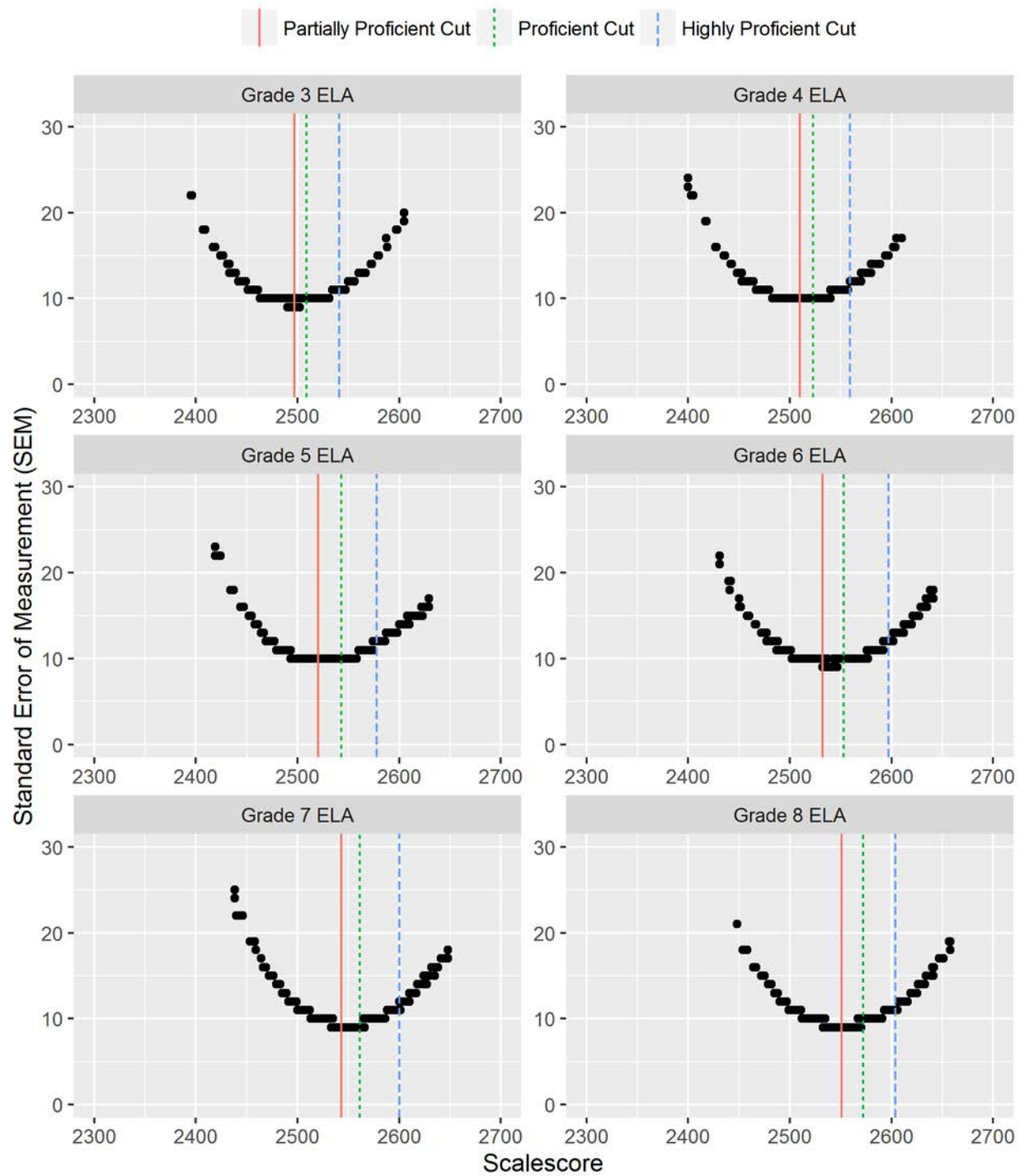
4.6.2 STANDARD ERROR OF MEASUREMENT

Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low and high performing students, the precision of test scores decreases near the tails of the ability distribution.

The figures in Exhibit 4.6.2.1 and Exhibit 4.6.2.2 present graphically the standard errors of measurement for the AzMERIT ELA and math assessments. Each figure also includes the location of the three AzMERIT performance standard cuts. As the figures indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Partially Proficient and Proficient performance standard cuts.²⁷ Test scores near the tails of the ability distribution are somewhat less precise, as expected. An SEM of .3 on the theta metric is consistent with an internal consistency of 0.9. The tables in Appendix I show the mean SEMs for students scoring in each of the performance levels on the AzMERIT reporting scale. While these tables and graphs indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance level classifications.

²⁷ Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 4.6.2.1 Overall Standard Error of Measurement for ELA



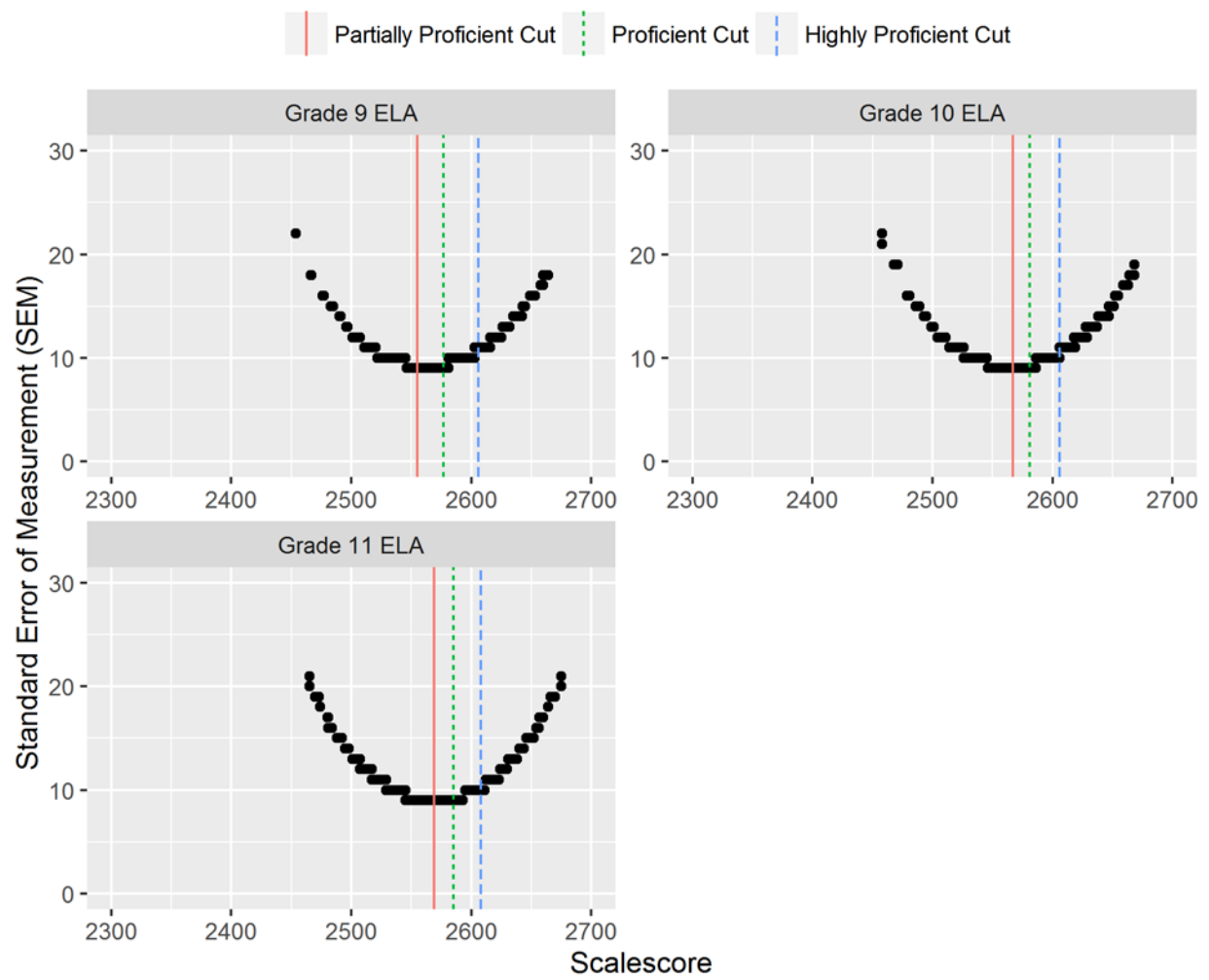
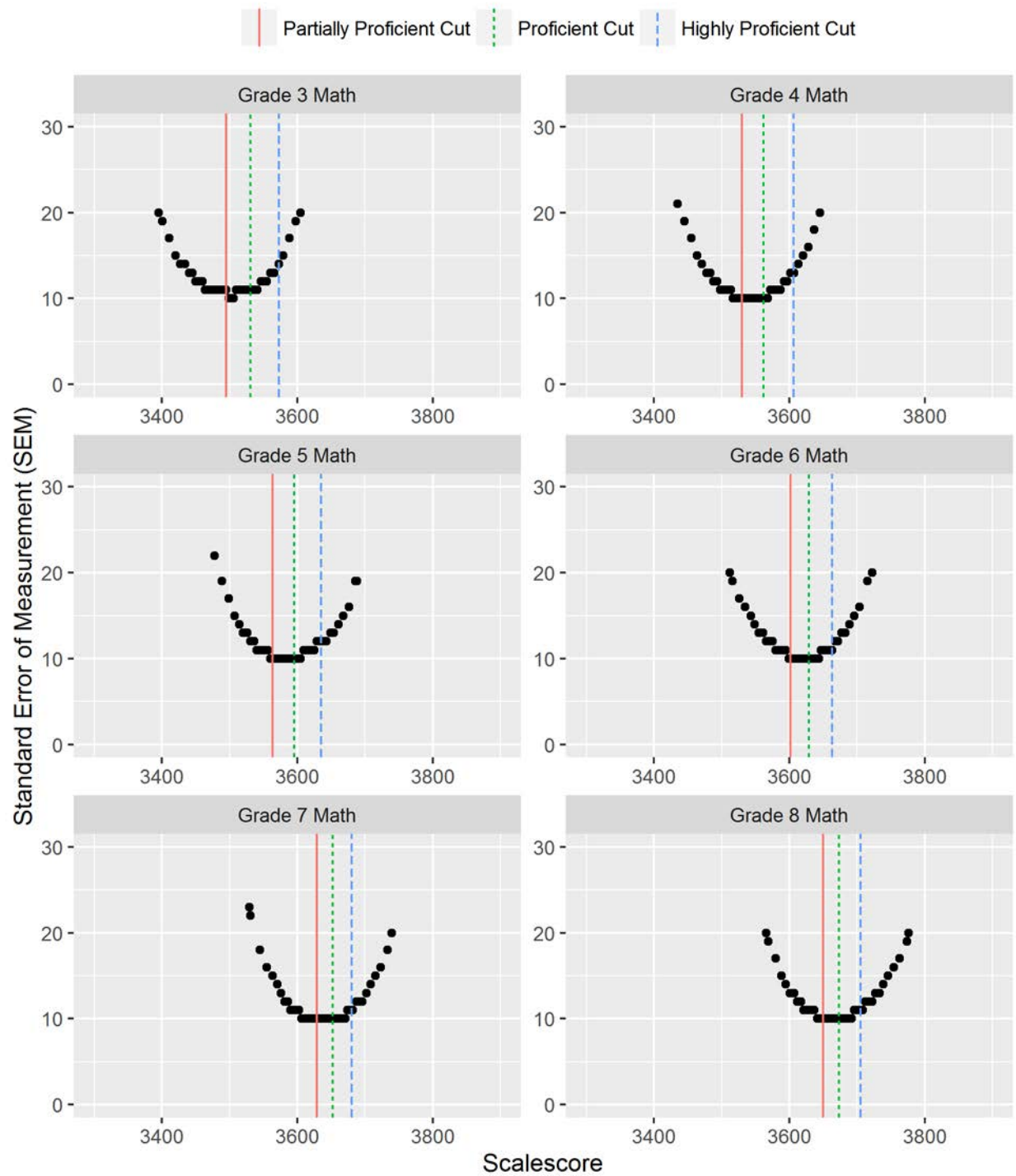
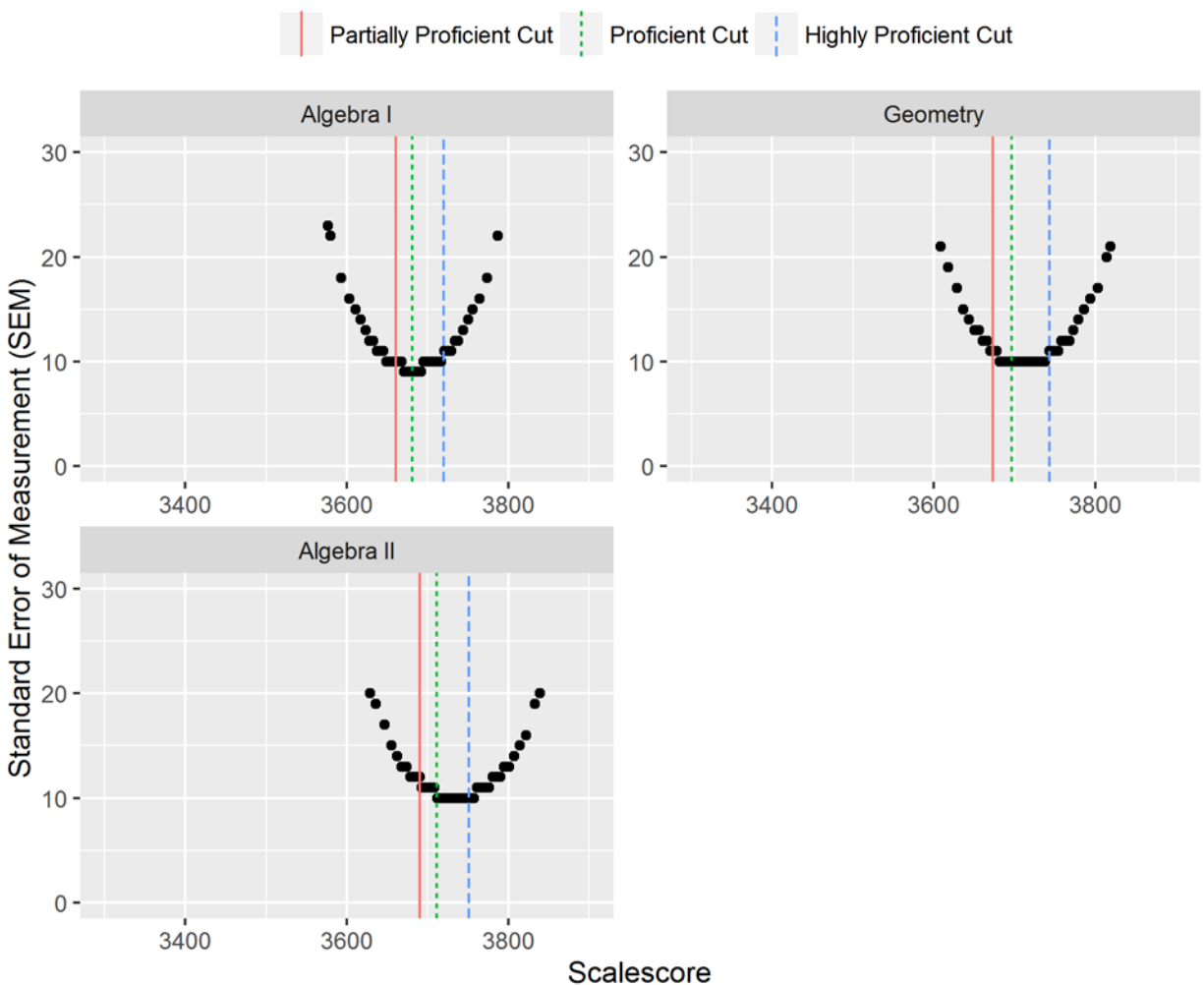


Exhibit 4.6.2.2 Overall Standard Error of Measurement for Math





4.6.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed to estimate the likelihood of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).²⁸ This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications is typically estimated on test scores based on a single test form from a single test administration using the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for classification accuracy and classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications

²⁸ Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

4.6.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the above probabilities directly using the likelihood function. The likelihood function of θ given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated ability (theta) is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 4.6.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications are different from students' observed scores (e.g., a student whose true score results in a classification as Proficient, but whose observed score results in an incorrect classification as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 4.6.4.1 Classification Accuracy

		Classification on the Form Actually Taken	
		Above the Cut Score	Below the Cut Score
Classification on True Score	At or Above the Cut Score	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

4.6.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 4.6.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 4.6.5.1 Classification Consistency

Classification on the 1st Form Taken		Classification on the 2nd Form Taken	
		Above the Cut Score	Below the Cut Score
		N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	At or Above the Cut Score		
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

4.6.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

Exhibit 4.6.6.1 presents the classification accuracy and consistency indexes for spring 2016 administration of AzMERIT. Exhibit 4.6.6.2 presents the classification accuracy and consistency indexes for each of the identified subgroups: gender (females and males), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with special education, free or reduced lunch, and accommodations). Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency index assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score plus the true score, which does not include measurement error.

Exhibit 4.6.6.1 Classification Accuracy and Consistency Estimates for Performance Standards Overall

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
3	0.91	0.91	0.95	0.88	0.88	0.93
4	0.92	0.91	0.95	0.88	0.88	0.93
5	0.92	0.92	0.94	0.89	0.88	0.92
6	0.92	0.92	0.97	0.89	0.89	0.96
7	0.92	0.91	0.96	0.88	0.88	0.95
8	0.92	0.92	0.96	0.88	0.89	0.95
9	0.91	0.91	0.96	0.88	0.88	0.95
10	0.90	0.92	0.96	0.87	0.89	0.95
11	0.91	0.93	0.96	0.88	0.90	0.95
MATH						
3	0.94	0.92	0.94	0.92	0.89	0.92
4	0.93	0.92	0.95	0.91	0.89	0.94
5	0.93	0.93	0.95	0.91	0.90	0.93
6	0.93	0.93	0.95	0.90	0.90	0.94
7	0.92	0.93	0.96	0.89	0.90	0.95
8	0.91	0.93	0.97	0.88	0.91	0.96
Algebra I	0.91	0.93	0.97	0.87	0.90	0.96
Geometry	0.89	0.93	0.98	0.85	0.90	0.97
Algebra II	0.88	0.93	0.98	0.83	0.90	0.97

Exhibit 4.6.6.2 Classification Accuracy and Consistency Estimates for Performance Standards across Subgroups

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA							
G3E	Overall	0.91	0.91	0.95	0.88	0.88	0.93
	Female	0.91	0.91	0.95	0.87	0.88	0.93
	Male	0.91	0.92	0.96	0.88	0.89	0.94
	African American	0.90	0.92	0.97	0.87	0.89	0.95
	Hispanic/ Latino	0.90	0.92	0.97	0.87	0.89	0.96
	Asian	0.92	0.92	0.91	0.90	0.88	0.88
	White	0.92	0.91	0.93	0.89	0.87	0.90
	Native Hawaiian/Pacific Islander	0.91	0.90	0.96	0.88	0.87	0.94
	American Indian or Alaskan	0.90	0.93	0.98	0.87	0.90	0.97
	Multiple Ethnicities	0.91	0.91	0.94	0.88	0.87	0.91
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.99
	Special Education	0.94	0.96	0.98	0.92	0.94	0.98
	Free/Reduced Lunch	0.90	0.92	0.97	0.87	0.89	0.96
	Accommodations	0.93	0.96	0.99	0.91	0.94	0.99
G4E	Overall	0.92	0.91	0.95	0.88	0.88	0.93
	Female	0.92	0.91	0.94	0.88	0.88	0.92
	Male	0.92	0.91	0.96	0.88	0.88	0.94
	African American	0.91	0.92	0.97	0.88	0.89	0.96
	Hispanic/ Latino	0.91	0.91	0.97	0.87	0.88	0.96
	Asian	0.94	0.91	0.91	0.92	0.88	0.88
	White	0.93	0.91	0.92	0.90	0.87	0.89
	Native Hawaiian/Pacific Islander	0.91	0.90	0.95	0.87	0.86	0.93
	American Indian or Alaskan	0.90	0.93	0.98	0.86	0.90	0.97
	Multiple Ethnicities	0.92	0.90	0.93	0.88	0.87	0.91
	Limited English Proficiency	0.94	0.97	1.00	0.91	0.95	0.99
	Special Education	0.94	0.96	0.99	0.92	0.94	0.98
	Free/Reduced Lunch	0.91	0.91	0.97	0.87	0.88	0.96
	Accommodations	0.94	0.96	1.00	0.91	0.95	0.99
G5E	Overall	0.92	0.92	0.94	0.89	0.88	0.92
	Female	0.93	0.91	0.93	0.90	0.88	0.91
	Male	0.92	0.92	0.95	0.89	0.89	0.94
	African American	0.91	0.92	0.97	0.88	0.89	0.95
	Hispanic/ Latino	0.91	0.92	0.96	0.88	0.88	0.95
	Asian	0.95	0.92	0.90	0.93	0.89	0.86
	White	0.94	0.91	0.92	0.92	0.87	0.89
	Native Hawaiian/Pacific Islander	0.92	0.91	0.92	0.89	0.88	0.90
	American Indian or Alaskan	0.90	0.93	0.98	0.86	0.91	0.98
	Multiple Ethnicities	0.93	0.91	0.93	0.90	0.88	0.90
	Limited English Proficiency	0.92	0.97	1.00	0.89	0.96	0.99
	Special Education	0.94	0.97	0.99	0.91	0.95	0.98
	Free/Reduced Lunch	0.91	0.92	0.96	0.88	0.89	0.95
	Accommodations	0.93	0.97	1.00	0.90	0.96	0.99
	Overall	0.92	0.92	0.97	0.89	0.89	0.96
	Female	0.92	0.92	0.97	0.89	0.89	0.95
	Male	0.92	0.92	0.98	0.89	0.90	0.97

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G6E	African American	0.91	0.93	0.99	0.88	0.91	0.98
	Hispanic/ Latino	0.91	0.93	0.99	0.87	0.90	0.98
	Asian	0.94	0.92	0.94	0.92	0.89	0.91
	White	0.93	0.91	0.95	0.90	0.87	0.93
	Native Hawaiian/Pacific Islander	0.91	0.91	0.96	0.87	0.88	0.95
	American Indian or Alaskan	0.91	0.95	0.99	0.88	0.93	0.99
	Multiple Ethnicities	0.92	0.91	0.97	0.89	0.88	0.95
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.98	1.00
	Special Education	0.95	0.98	1.00	0.93	0.96	0.99
	Free/Reduced Lunch	0.91	0.93	0.99	0.87	0.90	0.98
	Accommodations	0.95	0.98	1.00	0.92	0.98	1.00
G7E	Overall	0.92	0.91	0.96	0.88	0.88	0.95
	Female	0.91	0.91	0.96	0.88	0.87	0.94
	Male	0.92	0.92	0.97	0.88	0.89	0.96
	African American	0.91	0.91	0.98	0.87	0.88	0.97
	Hispanic/ Latino	0.90	0.92	0.98	0.87	0.89	0.97
	Asian	0.94	0.90	0.94	0.92	0.87	0.91
	White	0.93	0.90	0.94	0.90	0.86	0.92
	Native Hawaiian/Pacific Islander	0.90	0.90	0.98	0.86	0.87	0.97
	American Indian or Alaskan	0.90	0.94	0.99	0.86	0.92	0.99
	Multiple Ethnicities	0.93	0.90	0.96	0.90	0.87	0.94
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.97	1.00
	Special Education	0.94	0.97	1.00	0.91	0.96	0.99
	Free/Reduced Lunch	0.90	0.92	0.98	0.87	0.89	0.97
	Accommodations	0.94	0.98	1.00	0.92	0.97	1.00
G8E	Overall	0.92	0.92	0.96	0.88	0.89	0.95
	Female	0.91	0.92	0.96	0.88	0.88	0.94
	Male	0.92	0.93	0.97	0.89	0.90	0.96
	African American	0.91	0.93	0.98	0.88	0.91	0.97
	Hispanic/ Latino	0.91	0.93	0.98	0.87	0.91	0.97
	Asian	0.94	0.91	0.92	0.91	0.88	0.89
	White	0.92	0.91	0.94	0.89	0.87	0.92
	Native Hawaiian/Pacific Islander	0.93	0.90	0.96	0.89	0.87	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.92	0.91	0.96	0.88	0.88	0.94
	Limited English Proficiency	0.96	0.98	0.99	0.94	0.97	0.99
	Special Education	0.96	0.98	1.00	0.94	0.97	0.99
	Free/Reduced Lunch	0.91	0.93	0.98	0.87	0.91	0.97
	Accommodations	0.96	0.99	1.00	0.95	0.98	1.00
G9E	Overall	0.91	0.91	0.96	0.88	0.88	0.95
	Female	0.91	0.91	0.96	0.87	0.87	0.94
	Male	0.91	0.92	0.97	0.88	0.89	0.95
	African American	0.91	0.93	0.98	0.87	0.90	0.97
	Hispanic/ Latino	0.90	0.93	0.98	0.86	0.90	0.97
	Asian	0.93	0.90	0.93	0.91	0.86	0.90
	White	0.92	0.90	0.94	0.89	0.86	0.92
	Native Hawaiian/Pacific Islander	0.90	0.90	0.97	0.87	0.86	0.96

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G10E	American Indian or Alaskan	0.89	0.95	0.99	0.85	0.93	0.98
	Multiple Ethnicities	0.92	0.91	0.96	0.89	0.87	0.94
	Limited English Proficiency	0.93	0.96	0.99	0.90	0.94	0.98
	Special Education	0.94	0.97	0.99	0.91	0.96	0.99
	Free/Reduced Lunch	0.90	0.93	0.98	0.87	0.90	0.97
	Accommodations	0.95	0.98	1.00	0.93	0.97	1.00
	Overall	0.90	0.92	0.96	0.87	0.89	0.95
	Female	0.90	0.92	0.96	0.86	0.89	0.94
	Male	0.91	0.93	0.97	0.88	0.90	0.96
	African American	0.91	0.94	0.98	0.88	0.91	0.97
	Hispanic/ Latino	0.91	0.94	0.98	0.87	0.91	0.97
	Asian	0.91	0.91	0.93	0.88	0.87	0.90
	White	0.90	0.90	0.95	0.86	0.87	0.93
	Native Hawaiian/Pacific Islander	0.91	0.93	0.97	0.88	0.90	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.91	0.92	0.95	0.87	0.89	0.94
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.98
	Special Education	0.95	0.98	0.99	0.93	0.97	0.99
	Free/Reduced Lunch	0.91	0.94	0.98	0.88	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.95	0.98	0.99
G11E	Overall	0.91	0.93	0.96	0.88	0.90	0.95
	Female	0.91	0.92	0.96	0.87	0.89	0.94
	Male	0.92	0.94	0.97	0.89	0.91	0.95
	African American	0.92	0.95	0.98	0.89	0.93	0.97
	Hispanic/ Latino	0.92	0.94	0.98	0.89	0.92	0.97
	Asian	0.92	0.91	0.93	0.89	0.88	0.90
	White	0.91	0.91	0.94	0.87	0.88	0.92
	Native Hawaiian/Pacific Islander	0.93	0.95	0.96	0.91	0.93	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.92	0.92	0.95	0.89	0.89	0.93
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.98
	Special Education	0.96	0.98	0.99	0.95	0.97	0.99
	Free/Reduced Lunch	0.92	0.95	0.98	0.89	0.93	0.97
	Accommodations	0.97	0.98	1.00	0.96	0.98	0.99
MATH							
G3M	Overall	0.94	0.92	0.94	0.92	0.89	0.92
	Female	0.94	0.92	0.94	0.92	0.89	0.92
	Male	0.95	0.93	0.94	0.92	0.90	0.92
	African American	0.93	0.93	0.96	0.91	0.90	0.95
	Hispanic/ Latino	0.93	0.92	0.96	0.90	0.89	0.94
	Asian	0.97	0.94	0.90	0.96	0.91	0.86
	White	0.96	0.92	0.92	0.94	0.89	0.88
	Native Hawaiian/Pacific Islander	0.94	0.92	0.93	0.91	0.89	0.91
	American Indian or Alaskan	0.91	0.93	0.97	0.88	0.90	0.96
	Multiple Ethnicities	0.95	0.92	0.93	0.93	0.89	0.90
	Limited English Proficiency	0.92	0.95	0.99	0.88	0.93	0.98
	Special Education						

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G4M	Special Education	0.93	0.96	0.98	0.91	0.94	0.97
	Free/Reduced Lunch	0.93	0.92	0.96	0.90	0.89	0.94
	Accommodations	0.92	0.95	0.99	0.89	0.93	0.98
	Overall	0.93	0.92	0.95	0.91	0.89	0.94
	Female	0.93	0.92	0.96	0.90	0.89	0.94
	Male	0.94	0.92	0.95	0.91	0.90	0.94
	African American	0.92	0.93	0.98	0.89	0.91	0.97
	Hispanic/ Latino	0.92	0.92	0.97	0.89	0.90	0.96
	Asian	0.97	0.93	0.90	0.96	0.90	0.87
	White	0.95	0.92	0.93	0.93	0.89	0.90
	Native Hawaiian/Pacific Islander	0.95	0.91	0.96	0.92	0.88	0.94
	American Indian or Alaskan	0.91	0.93	0.98	0.87	0.91	0.98
	Multiple Ethnicities	0.94	0.92	0.94	0.91	0.89	0.92
	Limited English Proficiency	0.91	0.96	0.99	0.87	0.95	0.99
	Special Education	0.93	0.96	0.99	0.90	0.95	0.98
	Free/Reduced Lunch	0.92	0.92	0.97	0.89	0.90	0.96
	Accommodations	0.92	0.96	0.99	0.88	0.95	0.99
	Overall	0.93	0.93	0.95	0.91	0.90	0.93
	Female	0.93	0.92	0.95	0.90	0.89	0.93
G5M	Male	0.93	0.93	0.95	0.91	0.90	0.93
	African American	0.92	0.93	0.97	0.88	0.91	0.96
	Hispanic/ Latino	0.92	0.93	0.96	0.89	0.90	0.95
	Asian	0.97	0.94	0.92	0.95	0.91	0.88
	White	0.95	0.92	0.93	0.93	0.89	0.91
	Native Hawaiian/Pacific Islander	0.93	0.94	0.93	0.91	0.91	0.91
	American Indian or Alaskan	0.91	0.93	0.98	0.87	0.91	0.97
	Multiple Ethnicities	0.94	0.92	0.94	0.92	0.89	0.92
	Limited English Proficiency	0.91	0.96	0.99	0.88	0.95	0.99
	Special Education	0.93	0.97	0.99	0.90	0.95	0.98
	Free/Reduced Lunch	0.92	0.93	0.97	0.89	0.90	0.95
	Accommodations	0.92	0.97	0.99	0.88	0.95	0.99
	Overall	0.93	0.93	0.95	0.90	0.90	0.94
	Female	0.93	0.93	0.95	0.90	0.90	0.93
	Male	0.93	0.93	0.96	0.90	0.91	0.94
	African American	0.92	0.95	0.97	0.89	0.92	0.96
	Hispanic/ Latino	0.92	0.94	0.97	0.89	0.91	0.96
	Asian	0.96	0.93	0.93	0.94	0.91	0.90
	White	0.94	0.92	0.93	0.92	0.89	0.91
G6M	Native Hawaiian/Pacific Islander	0.92	0.93	0.95	0.89	0.90	0.93
	American Indian or Alaskan	0.91	0.95	0.98	0.88	0.92	0.97
	Multiple Ethnicities	0.94	0.92	0.95	0.91	0.89	0.93
	Limited English Proficiency	0.94	0.97	0.99	0.91	0.96	0.99
	Special Education	0.95	0.97	0.99	0.93	0.96	0.99
	Free/Reduced Lunch	0.92	0.94	0.97	0.89	0.91	0.96
	Accommodations	0.94	0.97	0.99	0.92	0.96	0.99
	Overall	0.92	0.93	0.96	0.89	0.90	0.95
	Female	0.92	0.92	0.96	0.88	0.90	0.95
	Male	0.92	0.93	0.96	0.89	0.91	0.95
	African American	0.92	0.95	0.98	0.88	0.93	0.97

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G7M	Hispanic/ Latino	0.91	0.94	0.98	0.88	0.92	0.96
	Asian	0.94	0.91	0.94	0.92	0.88	0.91
	White	0.92	0.91	0.94	0.89	0.88	0.92
	Native Hawaiian/Pacific Islander	0.90	0.93	0.98	0.86	0.91	0.97
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.98
	Multiple Ethnicities	0.92	0.92	0.96	0.88	0.89	0.94
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.98	0.99
	Special Education	0.96	0.98	0.99	0.94	0.97	0.99
	FRL	0.91	0.94	0.98	0.88	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.94	0.98	0.99
G8M	Overall	0.91	0.93	0.97	0.88	0.91	0.96
	Female	0.91	0.93	0.97	0.87	0.90	0.95
	Male	0.92	0.94	0.97	0.88	0.91	0.96
	African American	0.91	0.95	0.98	0.88	0.93	0.97
	Hispanic/ Latino	0.91	0.94	0.98	0.88	0.92	0.97
	Asian	0.93	0.92	0.93	0.90	0.88	0.91
	White	0.91	0.92	0.95	0.88	0.88	0.94
	Native Hawaiian/Pacific Islander	0.91	0.93	0.95	0.88	0.90	0.94
	American Indian or Alaskan	0.92	0.96	0.99	0.88	0.94	0.98
	Multiple Ethnicities	0.91	0.94	0.96	0.87	0.91	0.95
	Limited English Proficiency	0.95	0.98	0.99	0.93	0.97	0.99
	Special Education	0.95	0.98	0.99	0.93	0.97	0.99
	Free/Reduced Lunch	0.91	0.94	0.98	0.87	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.94	0.98	0.99
Algebra I	Overall	0.91	0.93	0.97	0.87	0.90	0.96
	Female	0.91	0.92	0.97	0.87	0.90	0.96
	Male	0.91	0.94	0.97	0.87	0.91	0.96
	African American	0.90	0.94	0.99	0.86	0.91	0.98
	Hispanic/ Latino	0.90	0.94	0.98	0.86	0.91	0.97
	Asian	0.94	0.93	0.93	0.92	0.90	0.90
	White	0.92	0.92	0.95	0.88	0.89	0.94
	Native Hawaiian/Pacific Islander	0.88	0.91	0.98	0.84	0.88	0.97
	American Indian or Alaskan	0.90	0.95	0.99	0.86	0.94	0.99
	Multiple Ethnicities	0.91	0.92	0.97	0.87	0.89	0.95
	Limited English Proficiency	0.92	0.96	0.99	0.88	0.94	0.98
	Special Education	0.93	0.97	1.00	0.89	0.96	0.99
	Free/Reduced Lunch	0.90	0.94	0.98	0.86	0.92	0.98
	Accommodations	0.93	0.98	1.00	0.90	0.97	0.99
Geometry	Overall	0.89	0.93	0.98	0.85	0.90	0.97
	Female	0.89	0.92	0.98	0.85	0.89	0.97
	Male	0.90	0.93	0.98	0.86	0.91	0.97
	African American	0.88	0.94	0.99	0.84	0.91	0.99
	Hispanic/ Latino	0.88	0.94	0.99	0.83	0.91	0.99
	Asian	0.94	0.92	0.94	0.91	0.89	0.92
	White	0.91	0.92	0.97	0.87	0.89	0.95
	Native Hawaiian/Pacific Islander	0.90	0.93	0.97	0.86	0.91	0.96
	American Indian or Alaskan	0.87	0.94	0.99	0.82	0.92	0.99

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
Algebra II	Multiple Ethnicities	0.89	0.93	0.96	0.85	0.90	0.95
	Limited English Proficiency	0.89	0.95	0.99	0.84	0.93	0.99
	Special Education	0.90	0.97	1.00	0.86	0.96	0.99
	Free/Reduced Lunch	0.88	0.94	0.99	0.83	0.91	0.98
	Accommodations	0.90	0.98	1.00	0.86	0.97	1.00
	Overall	0.88	0.93	0.98	0.83	0.90	0.97
	Female	0.87	0.93	0.98	0.83	0.89	0.97
	Male	0.88	0.93	0.98	0.84	0.91	0.97
	African American	0.88	0.94	0.99	0.83	0.92	0.99
	Hispanic/ Latino	0.87	0.94	0.99	0.82	0.91	0.98
	Asian	0.91	0.93	0.95	0.88	0.89	0.93
	White	0.89	0.92	0.97	0.85	0.88	0.96
	Native Hawaiian/Pacific Islander	0.86	0.96	0.99	0.82	0.93	0.98
	American Indian or Alaskan	0.87	0.96	1.00	0.82	0.93	0.99
	Multiple Ethnicities	0.88	0.94	0.97	0.83	0.91	0.96
	Limited English Proficiency	0.88	0.94	0.99	0.83	0.91	0.98
	Special Education	0.90	0.97	1.00	0.86	0.96	0.99
	Free/Reduced Lunch	0.87	0.94	0.99	0.82	0.92	0.99
	Accommodations	0.91	0.98	1.00	0.87	0.97	1.00

4.6.7 RELIABILITY FOR SUB-GROUPS IN THE POPULATION

Exhibit 4.6.7.1 and 4.6.7.2 shows the mean reliability for each of the identified subgroups: gender (females and males), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with IEPs (Special Education)²⁹, free or reduced lunch, and accommodations). As the Exhibit indicates, internal consistency reliabilities are generally stable across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups, and that test scores are similarly precise across demographic subgroups. For subgroups where the reliability coefficients are attenuated, there is a corresponding decrease in the subgroup variance relative to the overall student population, indicating that attenuation of reliability in subgroups is due to a restriction of range.

²⁹ Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Exhibit 4.6.7.1 Internal Consistency Reliability by Subgroup– ELA

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodations
ELA															
3	Reliability	0.89	0.89	0.89	0.88	0.89	0.88	0.87	0.83	0.89	0.89	0.86	0.77	0.87	0.80
	Variance	999	982	999	866	1094	917	823	636	997	1023	827	487	817	552
4	Reliability	0.89	0.89	0.89	0.89	0.89	0.89	0.88	0.85	0.88	0.88	0.87	0.78	0.88	0.81
	Variance	1099	1080	1095	1014	1111	1000	948	749	1022	1002	918	532	932	632
5	Reliability	0.90	0.90	0.90	0.88	0.90	0.90	0.88	0.85	0.89	0.89	0.87	0.81	0.89	0.81
	Variance	1154	1111	1150	1064	1114	994	1020	798	1059	1109	909	608	1011	642
6	Reliability	0.90	0.90	0.90	0.88	0.90	0.90	0.88	0.85	0.89	0.90	0.83	0.74	0.88	0.75
	Variance	1089	1080	1073	919	1123	1060	903	712	1018	1057	697	449	896	469
7	Reliability	0.89	0.89	0.90	0.89	0.89	0.88	0.88	0.85	0.88	0.89	0.83	0.76	0.88	0.77
	Variance	973	916	999	898	999	799	844	676	901	933	676	502	823	509
8	Reliability	0.90	0.90	0.90	0.89	0.90	0.90	0.89	0.85	0.89	0.90	0.82	0.86	0.89	0.76
	Variance	1016	964	1028	941	1049	970	883	650	956	966	632	761	858	475
9	Reliability	0.89	0.88	0.89	0.88	0.89	0.88	0.88	0.84	0.89	0.89	0.83	0.87	0.87	0.77
	Variance	909	851	928	834	920	789	789	606	873	897	643	806	780	484
10	Reliability	0.87	0.87	0.88	0.86	0.88	0.89	0.85	0.80	0.87	0.88	0.80	0.84	0.85	0.74
	Variance	764	722	787	678	865	890	630	471	759	802	523	614	631	400
11	Reliability	0.89	0.56	0.90	0.88	0.90	0.89	0.87	0.83	0.90	0.90	0.81	0.86	0.87	0.78
	Variance	927	871	946	798	1003	860	761	563	929	1016	611	733	759	529

Exhibit 4.6.7.2 Internal Consistency Reliability by Subgroup – Math

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodations
Mathematics															
3	Reliability	0.91	0.91	0.92	0.92	0.88	0.91	0.91	0.90	0.89	0.91	0.92	0.90	0.91	0.91
	Variance	1711	1571	1844	1788	1468	1570	1621	1391	1506	1731	1954	1384	1612	1674
4	Reliability	0.92	0.91	0.92	0.91	0.89	0.91	0.91	0.90	0.90	0.91	0.91	0.87	0.91	0.89
	Variance	1518	1419	1613	1431	1405	1440	1345	1169	1385	1382	1466	994	1343	1171
5	Reliability	0.92	0.92	0.92	0.91	0.91	0.91	0.91	0.90	0.91	0.92	0.89	0.87	0.91	0.87
	Variance	1620	1515	1716	1450	1507	1379	1462	1234	1497	1551	1367	1058	1428	1135
6	Reliability	0.92	0.92	0.92	0.91	0.92	0.92	0.91	0.89	0.92	0.92	0.88	0.85	0.91	0.86
	Variance	1685	1564	1792	1493	1788	1735	1490	1257	1546	1623	1373	1068	1485	1186
7	Reliability	0.91	0.90	0.91	0.89	0.91	0.90	0.89	0.86	0.90	0.91	0.83	0.79	0.89	0.79
	Variance	1263	1187	1331	1069	1356	1194	1075	866	1229	1265	819	645	1073	672
8	Reliability	0.90	0.90	0.90	0.88	0.91	0.91	0.89	0.86	0.90	0.90	0.81	0.81	0.88	0.78
	Variance	1234	1170	1288	1072	1422	1328	1138	929	1189	1239	827	790	1101	737
AlgI	Reliability	0.90	0.89	0.90	0.86	0.91	0.86	0.87	0.82	0.91	0.90	0.75	0.84	0.86	0.73
	Variance	1126	1045	1198	822	1339	781	896	665	1198	1064	533	802	861	522
Geo	Reliability	0.88	0.88	0.89	0.83	0.92	0.89	0.83	0.78	0.90	0.89	0.72	0.81	0.83	0.64
	Variance	1216	1140	1290	906	1508	1227	877	705	1324	1215	646	858	898	543
AlgII	Reliability	0.85	0.84	0.86	0.77	0.91	0.82	0.78	0.69	0.87	0.87	0.65	0.80	0.76	0.56
	Variance	1004	926	1084	720	1433	863	734	545	1120	1118	551	810	690	464

4.6.8 SUBSCALE RELIABILITY

Coefficient alpha estimates of internal consistency reliability associated with the subscales for the 2016 operational forms are presented in Exhibit 4.6.8.1-4.6.8.6. As indicated in the Exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT.

Exhibit 4.6.8.1 Subscale Reliabilities – ELA Grades 3-11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 3	0.74	0.73	0.76
Grade 4	0.76	0.77	0.70
Grade 5	0.80	0.71	0.74
Grade 6	0.82	0.71	0.73
Grade 7	0.77	0.72	0.72
Grade 8	0.79	0.72	0.76
Grade 9	0.75	0.74	0.74
Grade 10	0.74	0.65	0.74
Grade 11	0.78	0.70	0.75

Exhibit 4.6.8.2 Subscale Reliabilities – Math Grades 3-5

	Numbers & Operations- Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
Grade 3	0.64	0.74	0.82
Grade 4	0.70	0.63	0.86
Grade 5	0.76	0.77	0.83

Exhibit 4.6.8.3 Subscale Reliabilities – Math Grades 6 & 7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
Grade 6	0.78	0.77	0.72	0.58
Grade 7	0.68	0.68	0.67	0.73

Exhibit 4.6.8.4 Subscale Reliabilities – Math Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability & the Number System
Grade 8	0.79	0.59	0.59	0.57

Exhibit 4.6.8.5 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.80	0.75	0.62
Algebra II	0.67	0.51	0.63

Exhibit 4.6.8.6 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.58	0.62	0.62	0.67

4.7 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 4.7.1-4.7.6. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.³⁰ The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y .

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Exhibit 4.7.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 3 to 11

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
3	Literature	0.71		0.97	
	Writing & Language	0.64	0.63	0.86	0.84
4	Literature	0.74		0.97	
	Writing & Language	0.64	0.63	0.87	0.86
5	Literature	0.72		0.96	
	Writing & Language	0.68	0.62	0.93	0.86
6	Literature	0.73		0.96	
	Writing & Language	0.67	0.61	0.93	0.85
7	Literature	0.72		0.97	
	Writing & Language	0.66	0.65	0.92	0.91
8	Literature	0.71		0.94	
	Writing & Language	0.68	0.62	0.92	0.83
9	Literature	0.71		0.95	
	Writing & Language	0.65	0.64	0.87	0.86
10	Literature	0.66		0.95	
	Writing & Language	0.63	0.58	0.91	0.83
11	Literature	0.69		0.94	
	Writing & Language	0.65	0.62	0.90	0.85

³⁰ Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.

Exhibit 4.7.2 Subscale Intercorrelations– Math Grade 3 to 5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.68		0.99	
	OAT_NBT	0.72	0.77	0.98	0.98
4	MDG	0.69		1.00	
	OAT_NBT	0.77	0.76	0.99	0.99
5	MDG	0.73		0.95	
	OAT_NBT	0.77	0.77	0.96	0.96

Note: NF = Numbers and Operations-Fractions; MDG = Measurement, Data & Geometry; OAT_NBT = Operations and Algebraic Thinking, and Numbers in Base Ten.

Exhibit 4.7.3 Subscale Intercorrelations– Math Grade 6 & 7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.78			1.00		
	RP	0.78	0.78		1.00	1.00	
	GSP	0.68	0.68	0.68	1.00	1.00	1.00
7	NS	0.70			1.00		
	RP	0.72	0.71		1.00	1.00	
	GSP	0.69	0.71	0.71	0.99	1.00	1.00

Note: EE = Expressions and Equations; NS = Number System; RP = Ratio and Proportional Relationships; GSP = Geometry, Statistics and Probability.

Exhibit 4.7.4 Subscale Intercorrelations– Math Grade 8

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
8	Functions (F)	0.67			0.99		
	Geometry(G)	0.71	0.59		1.00	1.00	
	SPNS	0.71	0.60	0.62	1.00	1.00	1.00

Note: EE = Expressions and Equations; F = Functions; G = Geometry; SPNS = Statistics and Probability and the Number System.

Exhibit 4.7.5 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.76		0.98	
	Statistics	0.72	0.72	1.00	1.00
Algebra II	Functions	0.67		1.00	
	Statistics	0.66	0.61	1.00	1.00

Exhibit 4.7.6 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
-------	----------	-----------------------	--	----------------------------	--

		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.69			1.00		
	Modeling with Geometry (MG)	0.64	0.60		1.00	0.97	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.71	0.70	0.62	1.00	1.00	0.96

Note: CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations;

4.8 RATER EFFECTS

For grades in which statistical models were constructed for machine scoring of essay responses, Measurement, Inc. (MI) hand-scored over 4,100 responses per prompt, with each response double scored and any discrepant scores routed for a final resolution score. At each grade, students responded to one of six randomly selected writing tasks. Appendix J shows the rater agreement for each of the writing prompts per grade administered on the AzMERIT. The rater agreement reports show percentages of exact agreement (Equal), adjacent scores (Adj. Low or Adj. High) and nonadjacent scores (Low or High). The tables also identify mismatched scores when there is a difference involving nonscorable condition codes (Mismatch CC), or a nonscorable/scorable mix (MM CC/Score). Exhibit 4.8.1 provides a summary of those results, showing the mean exact agreement rate for dimension scores across grades. Generally exact agreement rates ranged from 65%-70%, with little variability across the essay prompts.

Exhibit 4.8.1 Mean Exact Agreement Rates for Online Essay Responses.

Dimension	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 11	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Purpose/Organization	64	2	72	2	70	1	63	2	70	3	65	1
Evidence/Elaboration	63	2	76	2	69	3	66	3	72	3	62	1
Conventions	70	2	66	1	73	1	67	3	73	4	74	2

For all grades, approximately 15% of essay responses for paper-based test administrations were double-scored. As the tables in Appendix J show, agreement rates for scores assigned to hand written responses were higher than achieved for responses made online. One possible reason for this effect is that paper scoring was conducted following scoring of online responses so that by the time of paper scoring, readers were better synced.

5. ITEM DEVELOPMENT & TEST CONSTRUCTION

The AzMERIT assessments are rigorously examined in accordance to the guidelines provided in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence based on these standards that is necessary to validate assessments for their intended purposes.

The AzMERIT assessments were designed to measure student progress toward achievement of the Arizona College and Career Ready Standards (ACCRS). Although the validity of AzMERIT test score interpretations are evaluated along several dimensions, as a criterion-referenced system of tests, the meaning of test scores are critically evaluated by the degree to which test content was aligned with the ACCRS.³¹

Alignment of content standards is achieved through a rigorous test development process that proceeded from the content standards and refers back to those standards in a highly iterative test development process that included the ADE, test developers, and educator and stakeholder committees. Items used to develop the spring 2016 operational test forms were drawn mainly from the AIRCore pool of items developed to align with the Common Core State Standards. These items were all reviewed by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. Only items that were found to align well with the ACCRS were used. To supplement the AIRCore pool of items, a few previously developed Arizona items that also aligned to the ACCRS were used.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards that are covered in each test administration. Thus, the test specification blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determined how student achievement of the ACCRS was evaluated, alignment of test blueprints with the content standards was critical. The ELA and math blueprints are also provided as an attachment in Appendix A.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet blueprint. Developing test forms is difficult because test blueprints could be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. In addition to meeting complex blueprint requirements, test developers worked to meet psychometric goals so that alternate test forms measure equivalently across the range of student ability.

In addition to these review-intensive item development and form construction processes, that ensures test forms meet complex blueprint specifications, Student Achievement Partners reviewed the AzMERIT English Language Arts and Math tests. The goal of their review was to determine how well the 2015 tests aligned to the Arizona

³¹ Standard 1.11 – When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

College and Career Ready Standards. This review was based on the criteria recently published in Criteria for Procuring and Evaluating High-Quality Assessments (Council of Chief State School Officers, 2014).

5.1 ITEM DEVELOPMENT PROCESS³²

The content development process for AzMERIT is managed within AIR's Item Tracking System (ITS), which acts as a content development and management tool, item bank, and publication system supporting both paper and online publication. This item development workflow leads items from inception, through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes and rationales at each review and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence, and that the review is conducted (or recorded in the case of committee review) by an authorized person. As items travel through Arizona's extensive review process, every version of every item is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

ITS allows remote Internet access by item writers and reviewers while ensuring security with individualized passwords for all users, limited access for external users, and strong encryption of all information. Upon publication, ITS tracks the item's use on test forms. After items are used, ITS stores the resulting statistics, including exposure statistics, classical item statistics, and statistics based on item response theory (IRT).

The AzMERIT item development process is predicated on a high level of interaction between test developers at AIR and ADE, as well as with Arizona educators and stakeholders. AIR's ITS manages item content throughout the entire life cycle of an item, from inception, through series of agreed-upon item review levels culminating in operational pool approval. It also manages item content beyond the operational life of the item, including migration of items for use in practice tests or other training materials. ITS ensures that every item follows through the entire sequence of development and provides Arizona and AIR management on-demand reports of the content and status of the inventory of items. Each item is directed through a sequence of reviews and sign-offs by AIR and ADE staff before it is locked for field test or operational administration.

The ITS is integrated with the item display engine used by the AzMERIT online test delivery system. This feature, combined with a "web approval" process, allows the display of online items to be "locked" well before test forms are constructed and ensures that only approved items are administered to Arizona students.

5.1.1 ITEM WRITING

Test development experts use item [specifications](#) to guide the item development process.³³ These item specifications, developed by content experts at AIR and ADE, strategically guide the item development process. They are detailed documents that specify content limits, model tasks, and response types for a particular standard. Item writers use these specifications while developing items to make the best use of the available item types.

³² Standard 4.7 – The procedures used to develop, review, and tryout items and to select items from the item pool should be documented.

³³ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

The item specifications were developed using a vertical alignment for each standard, wherein the suggested task demands and cognitive complexity of items build upon those of the previous grade level, just as the standards themselves do.

Additionally, the item specifications provide models for item writers. The models include item samples that target different Depth of Knowledge (DOK) and difficulty levels. These item models also annotate the information in order to communicate the intent of the standard and DOK and to clarify for the writer how to manipulate the item difficulty while keeping the cognitive demands the same.

Detailed item specifications include the following:

- **Content Limits:** This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. For example, in grade 3, fraction denominators are limited to 2, 3, 4, 6, and 8.
- **Acceptable Response Mechanisms:** This section identifies the various ways in which students may respond to a prompt—e.g., multiple choice, graphic response, proposition response, equation response, multi-select.
- **Depth of Knowledge:** The task demands of each standard can be classified as DOK 1, DOK 2, DOK 3 and/or DOK 4.
- **Task Demands:** In this section, the standards are broken down into specific task demands aligned to the standard. In addition, each task demand is assigned appropriate response mechanism, DOK, and practice clusters specifically relevant to that particular task demand.
- **Examples and Sample Items:** In this section, sample items are delineated along with their corresponding expected difficulties (easy, medium, and hard). Notes for modifying the difficulty of each task demand are detailed with suggestions for the item writer. The suggestions for adapting the difficulty based on the task demands are research-based and have been reviewed by both content experts and a cognitive psychologist.

Item writers consistently followed the item specifications during the item development process. During each level of review, items were compared to the item specifications to ensure their alignment to the standard, grade-level appropriateness, and adherence to the content limits set forth in the item specifications.

Within each grade or course, all items are aligned according to DOK. Depth of Knowledge, or commonly DOK, refers to the cognitive complexity of the item and the cognitive demands on the student. Based on work done by Webb (2002), there are four levels of DOK:

- **DOK 1—Recall.** Students recall basic mathematical ideas, perform basic arithmetic operations using established algorithms, and identify examples of general math principles.
- **DOK 2—Skill/Concept.** Students apply their basic knowledge (DOK 1) and extend their thinking to problem solve, identify relationships, and draw conclusions.
- **DOK 3—Strategic Thinking.** Students go beyond basic problem solving (e.g., word problems) to extend their thinking to nonroutine problem solving, hypothesize, and critique arguments or problem solving strategies.
- **DOK 4—Extended Thinking.** At this highest level, students engage in extended problem-solving activities, which require integration of multiple standards. For example, students may engage in a performance task that includes a common stimulus and four to six associated items related to the stimulus.

Depending upon the subject area and grade or course assessment, the percentage of items and score points aligned to DOK 1, DOK 2, DOK 3, and DOK 4 vary. The percentage of test items aligned to each DOK level for each assessment is indicated in the test construction blueprint. Although the exact number of items on each form may vary, the test specifications ensure that students are administered a substantial proportion of items that assess higher-order thinking skills.

ELA

ELA item development often begins with development of reading passages. AzMERIT passages represent a variety of genres and topics. AIR's content experts develop informational texts from multiple content areas, such as history, science, and technical subjects. Literary texts represent authentic pieces from multiple genres, including stories, poetry, and drama. The ratio of informational to literary texts increases at each grade band with a greater percentage of informational texts in the upper grades. The AzMERIT utilizes both single passages as well as passage sets in which students are asked to synthesize information across texts.

To ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading, test developers adhere to detailed passage specifications. Content experts use passage complexity worksheets—based on the passage specifications—to perform an in-depth analysis of each passage. The passage specifications call for a close examination of both quantitative measures, such as word counts and Lexile readabilities, as well as qualitative measures, such as passage structure and levels of meaning, all of which are defined as important measures of text complexity.

AzMERIT's ELA assessments include extended writing tasks that provide students with meaningful contexts in which to construct their responses. Each writing-prompt presents students with a variety of stimuli (usually at least two to three per task) that serve as a springboard for an informed piece of writing. Students are given research articles, charts and graphs, and narratives to serve as the basis for their written response. Students can then use this information, along with their own reasoning, to formulate an essay that is not only a clear and coherent expression of their own thinking but that is also grounded in research and evidence. Each student is administered a single informative/explanatory or opinion/argumentative writing essay.

Informative/explanatory writing is focused on conveying information accurately. Informative writing seeks to enlighten the reader about processes or procedures, phenomena, states of affairs, and terminology. To produce this kind of writing, students draw from what they already know as well as from primary and secondary sources. Students develop a controlling idea and a primary focus as they relate facts, details, and examples.

Opinion/argumentative prompts ask students to analyze primary and secondary sources, make sound judgments, and present their opinions and arguments in a coherent way that weaves personal opinion with evidence from the texts. The stimuli present opposing points of view about a topic so that students have enough information to take a stand. The stimuli are followed by a prompt that asks students to write an opinion/argumentative essay. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support the ideas they present.

Writing prompts present students with two or three passage stimuli on a single topic from science, technical subjects, or social studies. The reading level of the stimulus does not exceed the easy Lexile range for the grade level to enable the students to attend to the content of the passages and not struggle over unfamiliar language and non-content-related vocabulary. Moreover, this helps ensure students are assessed on their writing skills and not their reading abilities.

The stimulus is followed by a prompt that asks students to write a short essay about the topic. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support their main ideas. For example, the prompt might require students to describe the steps in a process or describe problems that need to be solved.

MATH

Calculators are not allowed for assessments at grades 3–6, while students participating in high school assessments are allowed continual access to specific calculator functions. For the grades 7 and 8 assessments, where calculator usage is allowable for some item types, the test items are grouped into two segments, administered separately to students: calculator and no-calculator. The construct of the items dictate which section they are to be assessed in.

5.1.2 MACHINE-SCORED CONSTRUCTED-RESPONSE ITEM DEVELOPMENT TOOLS

AzMERIT includes a number of machine-scored constructed response (MSCR) items which leverage a sophisticated system that allows for a large variety of item types expecting varied student responses to be developed, and scored efficiently and economically.

Machine-Scored Constructed-Response (MSCR) item development tools put the power of both item and rubric creation into the hands of item writers, and allow reviewers to score possible responses to ensure the rubric is enacted correctly. For example, when administered a graphic-response item, students can respond by drawing, moving, arranging, or selecting graphic regions. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer. Test developers have flexibility in identifying features of student responses to score, which go beyond simple features (e.g., whether the correct object is put in the correct place) but can involve abstraction. For example, if a student is asked to design an experiment, the rubric can discern whether the objects representing the experimental variable actually vary across conditions or cover the range of inquiry, among other capabilities. These concepts are abstracted and many different responses may reflect those abstract features. This ability enables machine rubrics to “justify” the partial credit assigned in terms of the skills that particular response features exemplify.

In addition, throughout the item development and review process, test developers can mimic the many different possible student responses, and review how the rubric is applied to those responses. Test developers can test the scoring rubric and make corrections to the scoring logic at each step.

When creating equation items, test developers have access to the Equation Editor tool. Student responses can be simple numeric responses or complex equations or even sets of equations. This tool allows for multiple answers and the development of multistep items. Test developers can customize the equation palette to show the appropriate functions. Just as the key pad is customizable, the answer spaces are as well. Additional answer spaces can be added as needed by the item writer. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer.

Such tools are integrated into the ITS, providing test developers the power and flexibility to use technology to create sophisticated AzMERIT items.

5.1.3 ITEM TYPES

AzMERIT includes a wide variety of item types that are designed around a broad and growing catalog of response mechanisms. In addition to selected response items, which include traditional multiple choice and more advanced multi-select and two-part items, AzMERIT tests utilize items with the following response mechanisms:

- Graphic Response, which includes any item to which students respond by drawing, moving, arranging, or selecting graphic regions.
- Hot Text, in which students select or rearrange sentences or phrases in a passage.
- Equation Response, in which students respond by entering an equation or number.
- Word Builder, in which students respond by entering a single number or word.
- Proposition Response, in which students respond in one or more English language sentences, which may be scored by our proposition-scoring engine, human scored, or a mixture of both.
- Essay Response, in which the student response is a longer written response.

AzMERIT items use technology to measure deeper knowledge and application of knowledge in a more open ended way and to machine score many such items. All MSCR items administered in AzMERIT are accessible. There could be occasions where it is necessary to sacrifice accessibility for some population to measure a critical standard, but test development staff would need to carefully consider the measurement benefit before developing that item.

Where possible, MSCR items were rendered for administration on paper test forms, using the gridded response field in the scannable answer documents. Where equation and graphic response items could not be rendered to accept a gridded response on paper forms, responses were hand-scored. For other MSCR items that could not readily be rendered for paper test administration, the item was replaced by another item measuring the same content standard(s).

The graphic-response mechanism supports most of the typical technology-enhanced item types, including sorting, matching, hot-spot, and drag-and-drop. In addition, it supports items where students actually draw a machine-scorable response and respond by constructing complex, open-ended diagrams, as well as many other possibilities. Because they are uniformly derived from a single response mechanism, the manipulations and interactions are consistent across these technology-enhanced item types, eliminating one possible source of construct-irrelevant variance.

Hot-text items are effectively selected-response items, though in some cases the number of potential selections is quite large. These machine-scored items can have multiple correct answers and allow for very flexible student responses.

The equation response mechanism asks students to enter one or more numbers, expressions, or equations using a palette of symbols. Test developers can specify which symbols are available on an item-by-item basis, or ADE can choose to have the palette remain consistent across all of the items within a grade level.

The availability of tools organized around response mechanisms creates a very flexible capability for test developers to create authentic, challenging tasks.

5.2 ITEM REVIEW

This section describes the multi-step item review process, that items travel through, from inception, to several rounds of test developer, ADE, and educator review, to field testing and final review prior to inclusion on operational test forms.³⁴ Items used to develop the spring 2016 operational test forms were mainly drawn from the AIRCore pool of items developed to align with the Common Core State Standards. These items were also reviewed for alignment to the Arizona College and Career Ready Standards (ACCRS) by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning ADE, Arizona educator committees, and parent/community committees act as reviewers throughout the item development cycle.

The item review procedures used to develop and review AzMERIT test items are designed to ensure item accuracy and alignment with the intended ACCRS. Following a standard item review process, item reviews proceed initially through a series of internal reviews before items are eligible for review by ADE content experts. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through four internal review steps before it is eligible for review by ADE. Those steps include

- Preliminary review, conducted by a group of AIR content area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert.

At every stage of the item review process, beginning with preliminary review, AIR's test developers analyze each item to ensure that

- The item is well-aligned with the intended content standard
- The item conforms to the item specifications for the target being assessed
- The item is based on a quality idea (i.e. it assesses something worthwhile in a reasonable way);
- The item is properly aligned to a depth of knowledge (DOK) level;
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward
- Any accompanying graphic and stimulus materials are actually necessary to answer the question
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question
- For selected response items, the set of response options is succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; and all plausible, but with only one correct option
- There is no obvious or subtle cluing within the item

³⁴ Standard 4.8 – The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

- The score points for constructed-response items are clearly defined; and
- For machine-scored constructed-response (MSCR) items, the items score as intended at each score point in the rubric.

Based on their review of each item, the test developer can accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to ADE for their review. At this stage, items may be further revised based on any edits or changes requested by ADE, or rejected outright. Items passing through ADE review level then have to pass through a stakeholder review, in which a educators review each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- aligned to the Arizona content standards
- appropriate for the grade level
- accurate
- presented online in a way that is clear and appropriate
- free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics.

Items successfully passing through this committee review process were then presented to a parent/community review committee to ensure that test content met community standards. Items successfully passing through all review levels were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is, therefore, an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass in each stage of a two-stage review before being included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct and there are no other obvious problems with the items.

ADE content staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, ADE determined that certain flagged items must be rejected, or deemed the item eligible for inclusion in operational test administrations.

5.3 FIELD TESTING

To establish a pool of items for constructing future AzMERIT test forms, newly developed test items were embedded in the spring 2015 and spring 2016 AzMERIT test forms for field-testing. Embedding field-test items in operational assessments yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations. A number of factors that may influence item difficulty in the context of operational test administrations may be less relevant in stand-alone field-test contexts. For example, in a high-stakes test, such as high school end-of-course (EOC) exams where test performance may impact student grades, students may be motivated to expend greater effort to achieve maximum performance. Conversely, the high-stakes assessments may also be more likely to elicit anxiety in some students, thus impairing their performance on the tests. Even when assessments are low stakes for students, schools often work to convey to students the importance of statewide assessments in ways that are likely not done for independent field tests. While the impact of contextual factors may not be great, embedded field testing ensures that all aspects of the operational testing context influencing item difficulty are incorporated into the resulting item parameter estimates.

Embedded field-testing is especially useful in the context of a pre-equating model for scoring and reporting test results. Because the test administration context remains the same between the embedded field test (EFT) and subsequent operational test administration, item parameter estimates are more stable over time than they may be when obtained through stand-alone field-testing.

A potential drawback of the EFT approach is the increased assessment burden placed on students and schools. For this reason, AzMERIT utilizes EFT designs for purposes of item bank maintenance. Arizona uses AIR's online field-test engine for computer-administered tests, which, when combined with Arizona's large student population, serves to greatly reduce the number of EFT slots necessary to replenish and even grow the item banks for the Arizona assessments.

The field test engine randomly samples field test items for each individual test administration, essentially creating thousands of unique EFT forms. This sampling approach to embedding field-test items results in several important outcomes:³⁵

- Reduction in the number of embedded field-test items that each student must respond to and more efficient "spiraling" of items, which reduces clustering of item responses, resulting in more precise parameter estimates
- More generalizable item statistics because they are not based on items appearing in a single position
- A truly representative sample of respondents for each item

The embedded field testing algorithm actually consists of two different algorithms – one for identifying which field test items will be administered to which student (the distribution algorithm), and one for selecting the position on the test for each item administered the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field test items designated for administration to each student. This

³⁵ Standard 4.9 – When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.

randomization ensures that a) each item is seen by a representative sample of Arizona students, and b) every item is as likely as every other item to appear in a class or school, minimizing clustering effects.

In addition, a fixed block of field test items was also embedded in paper AzMERIT test forms so that the number of items responded to by students did not vary between assessment modes.

Following the spring 2016 test administrations, the free calibration was performed on the operational items on each of the ELA and mathematics tests. Then the free calibrated item parameters were linked back to the 2015 spring scale using the mean-mean equating method. The tables in Appendix K present the linking constant, post-equated parameters and item drifts for each test. The field test item calibration was conducted by anchoring on the post-equated operational item parameters for all of the ELA and mathematics tests. However, only the ELA spring 2016 operational tests were scored using the post-equated item parameters.

5.4 ITEM STATISTICS

Following the close of test administration windows, AIR psychometrics staff worked to analyze field test data in preparation for item data review meetings and promotion of high quality test items to operational item pools.³⁶ Analysis of field test items includes classical item statistics as well as the IRT item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by AIR and ADE psychometric and content staff for possible miskey or scoring errors. Items that pass through AIR and ADE statistical review are accepted for future operational use. Appendix L provides the slide presentation used to train reviewers for item data review. The training is designed to ensure that all reviewers understand how items are evaluated and that they are interpreting item statistics correctly.

5.4.1 CLASSICAL STATISTICS

Classical item analyses ensured that the field test items function as intended with respect to the AzMERIT's underlying scales. AIR's analysis program computed the required item and test statistics for each multiple-choice and constructed-response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are either extremely difficult or extremely easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For multiple-choice items, the proportion of examinees in the sample selecting the correct answer (p -value) is computed, as well as those selecting the incorrect responses. For constructed-response items, item difficulty is calculated both as the item's mean score and as the average

³⁶ Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

proportion correct (analogous to p -value and indicating the ratio of an item's mean score divided by the number of points possible). Items are flagged for review if the p -value was less than .25 or greater than .95.

The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low- achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .25.

Distractor analysis for the multiple-choice items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response is greater than .05. In addition, items are flagged if the proportion of students responding to a distractor exceeds the proportion selecting the keyed response. Although non-modal response keys are typically observed with difficult items, in combination with poor item discrimination it may indicate a miskeyed item.

5.4.2 IRT STATS

Rasch and Masters' Partial Credit Model are used to estimate the item response theory (IRT) model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showing the item statistics resulting from the free (unanchored) estimation of parameters for items in the operational tests were reviewed, as well as the Winsteps-generated item and persons maps. Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are conservatively flagged if Infit or Outfit values are less than 0.7 or greater than 1.3.

5.4.3 ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, DIF can indicate bias, so all field-tested items were evaluated for DIF, and all items exhibiting DIF were flagged for further examination by AIR and ADE staff to make a final decision about whether the item should be excluded from the pool of potential items given its performance in field testing potential items.

AIR conducts DIF analysis on all field-tested items to detect potential item bias across major ethnic and gender groups. In Arizona, DIF is investigated among the following group comparisons (reference group/ focal group):

- Male/ Female
- Hispanic, Latino or Spanish origin/ Non-Hispanic
- White/ Black, African American, or Negro
- White/ American Indian or Alaskan Native
- White/ Asian
- White/ Native Hawaiian or Other Pacific Islander
- White/ Multiple ethnicities selected
- Special Education
- Limited English Proficiency
- Free or Reduced Lunch
- Accommodations

AIR uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. Because students within a district, school, and classroom are more similar than would be expected in a simple random sample of students statewide, the information provided by students within a school is not independent, so that standard errors based on the assumption of simple random samples are underestimated. We compute design consistent standard errors that reflect the clustered nature of educational systems. While clustering is mitigated through random administration of large numbers of embedded field test items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into a configurable number of intervals to compute the Mantel-Haenszel chi-square ($MH \chi^2$) DIF statistics. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta ($\Delta_{hat MH}$) for the MC items; the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the CR items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed below. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e., -A, -B, or -C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Exhibit 5.4.3.1. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or reference groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992, Camilli & Shepard, 1994, Muniz, Hambleton, & Xing, 2001, Sireci & Rios, 2013).

Exhibit 5.4.3.1 DIF Classification Rules

Item Type	Category	Rule
Dichotomous Items	C	$MH \chi^2$ is significant and $ \Delta_{hat MH} \geq 1.5$
	B	$MH \chi^2$ is significant and $ \Delta_{hat MH} < 1.5$
	A	$MH \chi^2$ is not significant.
Polytomous Items	C	$MH \chi^2$ is significant and $ SMD / SD \geq .25$
	B	$MH \chi^2$ is significant and $ SMD / SD < .25$
	A	$MH \chi^2$ is not significant.

5.5 TEST CONSTRUCTION

The process for constructing fixed-form operational tests begins after field testing and review of item performance. Once an operational item pool is established, AIR content specialists begin the process of constructing test forms. Operational passages and items qualified for operational forms are those that met all of the criteria established by ADE in terms of content, fairness review, and data characteristics.

5.5.1 OPERATIONAL FORM CONSTRUCTION

Each AzMERIT form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it is covered, the type of items that measure the constructs, and every other content-relevant aspect of the test. The statistical targets, which are held constant across years and across modes, ensure that students receive scores of similar precision, regardless of which form of the test they receive.³⁷

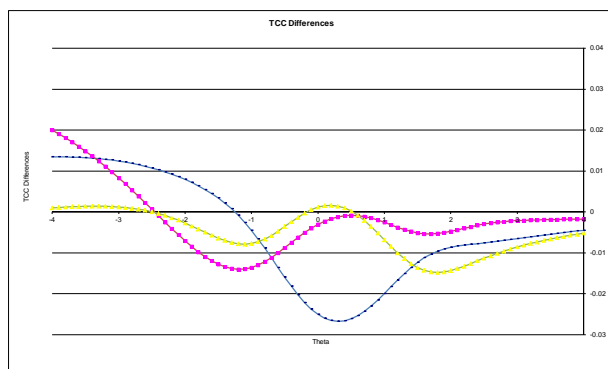
AIR's test developers used FormBuilder software to help construct operational forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively create test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers build a test map. This helps content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

Immediately upon generation of a test form, FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allow another opportunity to ensure that poorly performing items are not included in operational test forms.

As test developers built forms, the FormBuilder generated, TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 5.5.1.1 shows a sample graph of TCC differences. There are several important things to note when examining TCC differences. First, differences in TCCs can occur at specific locations in the TCCs across a range of abilities. These differences reflect different emphases in test information across forms at these ability levels. If the difficulty and error structure for the target forms is virtually identical to the reference form, as in the sample TCC and SEM curves, then the item selection process concludes with multiple, parallel test forms. Once the goal of parallel forms is achieved, the information is entered into ITS, which tracks item usage and generates bookmaps (test maps) for use in scoring, forms development, and other processes.

³⁷ Standard 4.12 – Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

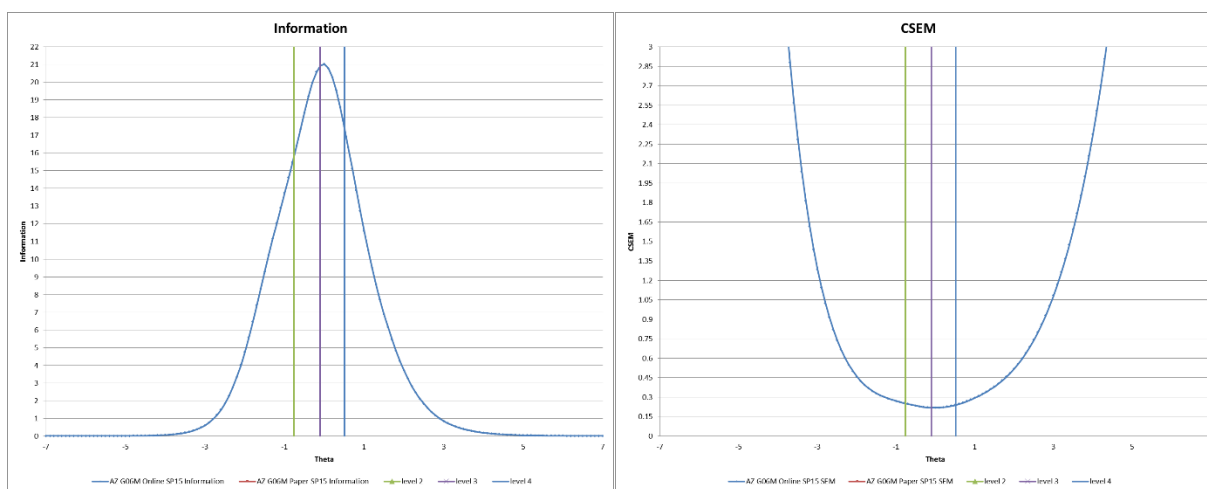
Exhibit 5.5.1.1 Test Characteristics Curve Differences



The reference form for each assessment is the operational test form administered in spring 2015. As illustrated in Exhibit 5.5.1.2, by evaluating test characteristics in reference to the base year forms, students are administered tests each year that are equivalent in difficulty across the range of ability. The Test Characteristic Curve (TCC) and SEM graphs that were used to evaluate the spring 2016 operational test forms are presented in Appendix M.

In addition, although paper test forms were developed to be as nearly identical to the online forms, there were some items that could not readily be rendered for paper test administration. In those instances, replacement items were identified and TCCs and SEMCs were evaluated to ensure equivalence between online and paper test forms.

Exhibit 5.5.1.2 Test Information and Standard Errors Relative to Performance Standards



5.5.2 ASSEMBLING TEST FORMS

The mechanical features of a test—arrangement, directions and production—are just as important as the quality of the items. Many factors directly affect a student’s ability to demonstrate proficiency on the assessment, while others relate to the ability to score the assessment accurately and efficiently. Still others affect the inferences made from the test results.

When the test developer reviews a test form for content, in addition to making sure all the benchmark/indicator item requirements are met, he or she also makes sure that the items on the form do not cue each other – that one item does not present material that indicates the answer to another item. This is important to ensure that a

student's response on any particular test item is unaffected by, and is statistically independent of, a response to any other test item. This is called "local independence." Independence is most commonly violated when there is a hint in one item about the answer to another item. In that case, a student's true ability on the second item is not being assessed.

Test Developers begin the form construction process by first identifying the pool of items from which forms are built. This pool of items resides at a locked operational status in the Item Tracking System. Each item contains a historical record that clearly demonstrates it has survived the full review process from internal development through client, committees, and its statistical data review.

Upon identifying and reviewing the eligible pool of items, a test developer then considers the limitations of the pool, if any. For example, there might be a shortage of high depth of knowledge (DOK 3) items at a particular benchmark. The test developer will review and select from among these items first to ensure that the constraints of the blueprint are met.

Once the items and passages for the form are selected and matched against the blueprint, the test developer reviews the form for a variety of additional content considerations, including the following:

- The items are sequentially ordered.
- Each item of the same type is presented in a consistent manner.
- The listing of the options for the multiple-choice items is consistent.
- The answer options are lettered with A, B, C, and D.
- All graphics are consistently presented.
- All tables and charts have titles and are consistently formatted.
- The number of the answer choice letters is approximately equal across the form.
- The answer key was checked by the initial reviewer and one additional independent reviewer.
- All stimuli have items associated with them.
- The topics of items, passages or stimuli are not too similar to one another.
- There are no errors in spelling, grammar or accuracy of graphics.
- The wording, layout and appearance of the item matches how the item was field-tested.
- There is gender and ethnic balance.
- The passage sets do not start with or end with a constructed response item.
- Each item and the form are checked against the appropriate style guide.
- The directions are consistent across items and were accurate.
- All copyrighted materials have up-to-date permissions agreements.
- Word counts are within documented ranges.

After completing the initial build of the form, the test developer hands it off to another content specialist, who conducts a final review of the criteria listed above. If the test specialist reviewer finds any issues, the form is sent back for revisions. If the form meets blueprint and complies with all specified criteria, the test developer sends it to the psychometric team for review. When the psychometric team approves the form, the test developer forwards the form evaluation workbooks to ADE for review and approval.

6. TEST ADMINISTRATION

6.1 ELIGIBILITY

Arizona public school students in Grade 3 and above were required to participate in AzMERIT testing.³⁸ Additionally, any student enrolled in a private school or Bureau of Indian Education school and any students who are home schooled had the option to participate as well. Students enrolled in Grades 3 – 8 took English Language Arts (ELA) and Math at the grade level in which they were enrolled. Students, in any grade, who are enrolled in high school level English language arts courses (Freshman English, Sophomore English, Junior English, or their equivalents) or high school level math courses (Algebra I, Geometry, Algebra II, or their equivalents) took the respective End-of-Course (EOC) test.

Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them eligible for the alternate assessment for ELA and Math were excluded from AzMERIT and instead took the Multi-State Alternate Assessment (MSAA).

6.2 ADMINISTRATION PROCEDURES

Key personnel involved with AzMERIT administration include the District Test Coordinators, School Test Coordinators, and Test Administrators who proctor the test. For information about the roles and responsibilities of testing staff, see below.

A secure browser developed by AIR was required to access the computer-based AzMERIT tests. The secure browser provided a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to desktop functionalities, such as the Internet and email. Other measures that protect the integrity and security of the online test are presented in “Test Security Procedures” below.

Prior to each test administration, statewide District Test Coordinator training sessions were conducted to provide information regarding both the paper and computer-based test administrations. The training also provided an overview of the Test Delivery System (TDS), Online Reporting System (ORS), and Test Information Distribution Engine (TIDE). Recorded training sessions and narrated training videos were posted online. The Test Coordinator Manual and Test Administration Directions were shipped to every testing district. Additionally, test administrators were required to complete the online TA Certification Course before administering a computer-based test.³⁹ District Test Coordinators and School Test Coordinators were responsible ensuring that all test administration personnel (paper and computer-based) were properly trained using the various resources prior to the start of testing.

³⁸ Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

³⁹ Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

Manuals and guides on test administrations are available on the AzMERIT Portal.⁴⁰ The Test Administrator User Guide was designed to familiarize Test Administrators with the Test Delivery System and contained tips and screenshots throughout the text. The guide provides enough how-to information to enable TAs to access and navigate the Test Delivery System. The user guide provides the following information:

- Steps to take prior to accessing the system and logging in
- Navigating the TA interface application
- The Student Interface, used by students for computer-based testing
- Training sites available for Test Administrators and students
- Secure browsers and keyboard shortcut keys

The *AzMERIT Test Coordinator's Manual* provides information about policies and procedures for AzMERIT Test Coordinators. This manual is updated prior to each test administration and includes test administration policies and guidance for Test Coordinators before, during, and after the window.

The *AzMERIT Test Administration Directions, End-of-Course* and the *AzMERIT Test Administration Directions, Grades 3-8* provide information about policies and procedures for the AzMERIT, both computer-based and paper-based versions. The *Test Administration Directions*, which is updated prior to each test administration, includes test administration information, guidance, and directions.

The *AzMERIT Test Administration Directions* provide easy-to-follow instructions for the online testing environment, such as creating online testing sessions, monitoring online sessions, verifying student information, assigning test accommodations, starting and pausing test sessions.⁴¹ Similar guidance is provided for the paper testing environment, including instructions for the paper testing session, monitoring sessions, verifying student information, and assigning test accommodations. Additional instructions for administering tests to students using Braille accommodated test booklets are provided in the *Supplemental Instructions for Braille* documents.

District and school personnel involved with AzMERIT test administration played an important role in ensuring the validity of the assessment by maintaining both standardized administration conditions and test security.

District Test Coordinators were responsible for coordinating testing at the district level. District Test Coordinators were ultimately accountable for ensuring that testing was conducted in accordance with the test security and other policies and procedures established by ADE. They ensured that the Test Administrators in each school were appropriately trained and aware of policies and procedures, and that they were trained to use the reporting system.

Districts may also identify School Test Coordinators. School Test Coordinators may assist in the identification and training of Test Administrators. They may also create testing schedules and procedures for the school. If the school administered AzMERIT online, the School Test Coordinators may work with Technology Coordinators to ensure that the necessary secure browsers were installed and any other technical issues were resolved. During the testing

⁴⁰ Standard 7.13 – Supporting documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to the appropriate people in a timely manner.

⁴¹ Standard 4.15 – The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.

window, School Test Coordinators needed to monitor testing progress, ensure that all students participate as appropriate, and handle testing incidents as necessary.

Test Administrators (TA) were responsible for reviewing necessary manuals and user guides to prepare the testing environment and ensuring that students did not have unapproved books, notes, or electronic devices out during testing. TAs were required to administer AzMERIT tests following the directions found in the *AzMERIT Test Administration Directions*.⁴² Any deviation in test administration must be reported by TAs to the School Test Coordinator, who reports it to the District Test Coordinator. The District Test Coordinator then reports it to ADE.

Test Administrators who administered computer-based AzMERIT tests conducted a training test session using the AzMERIT Sample Tests. Test Administrators were required to pass a qualifying test before they were eligible to administer the AzMERIT online.⁴³

Test Administrators must also ensure that only resources that were allowed for specific tests were available and no additional resources were being used during the test. No calculators were permitted in AzMERIT Math tests for grades 3-6. Scientific calculators were permitted in AzMERIT Math Part 1 for grades 7 and 8. Graphing calculators were permitted in AzMERIT Math EOC Parts 1 and 2 (Algebra I, Geometry, and Algebra II). Online calculators were provided as embedded tools within the appropriate computer-based test parts. Handheld calculators could be provided to students during the appropriate test sessions. Calculator guidance was provided in both the *AzMERIT Test Coordinator's Manual* and the *AzMERIT Test Administration Directions*. The online calculators were made publicly available on the AzMERIT Portal, as well as made securely available in a secure browser for paper-based test students to access, if needed. Providing a calculator with prohibited functionality or in the incorrect test session is cause for test invalidation.

For the computer-based ELA Reading tests, headphones or earbuds were required. There were no technical specifications for headphones or earbuds. The equipment was to be checked to ensure it worked with the computer or device the students would use for the assessment prior to the first day of testing. A sound test was also built in to the computer-based assessment and students were asked to verify that headphones and earbuds were working prior to entering the test.

For the paper-based AzMERIT tests, Test Administrators needed to ensure that students used No. 2 pencils to record their responses. School Test Coordinators provided TAs with the materials needed to administer each test session. Secure materials were delivered or picked up immediately before the beginning of each test session. During math testing and when responding to the writing prompt, students were permitted to use the scratch paper as a workspace. After testing, TAs needed to return the testing materials to the School Test Coordinator.

The School Test Coordinator and Test Administrators worked together to determine the most appropriate testing option(s) and testing environment and the average time needed to complete each test. The appropriate protocols were established to maintain a quiet testing environment throughout the testing session. TAs also needed to ensure that adequate time was available to start computers, load secure browsers, and log in students for computer-based tests and pass out and collect test booklets and materials for paper-based tests.

⁴² Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

⁴³ Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

MANAGING TESTING

To help schools manage their test schedule, allocate testing resources, and prioritize testing, the AzMERIT online reporting system, which is described in detail later in this chapter, offered participation reports for online testers. Within the online reporting system, educators can generate up-to-the-minute reports showing students' test status. In addition, users can set testing schedules, monitor testing progress across schools, and track students' participation based on their performance on previous tests.

ORS Online Reporting System

Logged in as: Doe, Jane | Contact Us | Log Out

AzMERIT

Home | Test Management Center

This page: ? Help Definitions

Plan and Manage Testing

Step 1: Choose What

Test:

Administration:

Test Name:

Enrolled Grade:

Step 2: Choose Who

District:

School:

Personnel:

Step 3: Get Specific

☒ Students who the test in the selected administration

☐ Students who have a status of in the selected administration

☐ Students whose most recent was between and

Note: If no TA or Session ID is specified, date range cannot exceed 15 days

☐ Search students by :

or

AzMERIT Help Desk
1.844.560.7812
azmerithelpdesk@air.org

6.3 TESTING CONDITIONS, TOOLS, AND ACCOMMODATIONS

This section summarizes the testing conditions, tools, and accommodations that are available to AzMERIT testers, as described in the *Testing Conditions, Tools, and Accommodations Guidance* manual that is available each administration. Test tools and accommodation requirements are designed to ensure that test content is accessible for all students.

6.3.1 UNIVERSAL TEST ADMINISTRATION CONDITIONS

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment.⁴⁴ Universal test administration conditions are available for both paper-based test (PBT) and computer-based testing (CBT) modes. Universal test administration conditions include:

⁴⁴ Standard 3.4 – Test takers should receive comparable treatment during the test administration and scoring process.

Standard 4.5 – If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- Extended time. (Testing session must be completed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan as an accommodation, for AzMERIT testing purposes these are not considered testing accommodations and are available to any student who needs them not just to students with IEPs.

6.3.2 UNIVERSAL TESTING TOOLS FOR COMPUTER BASED TESTERS

The AzMERIT computer-based testing platform offers numerous testing tools. All tools are available in the AzMERIT Sample Tests, which are available to test administrators and students prior to each test administration. Test administrators are encouraged to ensure that students who will participate in the computer-based AzMERIT take the AzMERIT Sample Tests and familiarize themselves with the available tools.

Exhibit 6.3.2.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Standard 6.4 – The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

Exhibit 6.3.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to click anywhere on the selected response text or button for multiple choice options.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen <i>Test Instructions and Help</i> .
Highlighter	Highlight text in a passage or item.
Line Reader	Allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	Allows the session to be paused at any time and restarted and taken over a one day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

6.3.3 SUBJECT AREA TOOLS FOR CBT AND PBT

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 6.3.3.1.

Exhibit 6.3.3.1 Subject Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	<p>CBT – Students have access to the dictionary/thesaurus tool. Students may opt to use a published, paper dictionary or thesaurus instead of using this tool.</p> <p>PBT – Schools must make published, paper dictionaries and thesauruses available to students.</p> <p>Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned-off.</p>
Writing Guide	Writing	<p>CBT – Students have access to the writing guide tool.</p> <p>PBT – The writing guide is included within the test booklet.</p>
Scratch Paper	Writing and Math	<p>CBT – Schools must provide scratch paper (plain, lined, or graph) to students</p> <p>PBT – Schools must provide scratch paper (plain, lined, or graph) to students</p>
Calculator Grades 7-8 (Part 1 only): scientific calculators are acceptable EOC (entire test): graphing calculators are acceptable	Math	<p>CBT – Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted.</p> <p>PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.</p>

6.3.4 ACCOMMODATIONS

Accommodations are provisions made in how a student accesses and demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Language Learner or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education need, or language need and the accommodation(s) provided to the student during educational activities, including assessment. Test administrators are instructed to make accommodation decisions based on individual needs, and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzMERIT test that is not already used regularly in the classroom.

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section.⁴⁵ Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student’s individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact ADE for guidance.

Allowable accommodations are described below.⁴⁶

ACCOMMODATIONS FOR STUDENTS WITH AN INJURY

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations described in Exhibit 6.3.4.1. There are no specific CBT tools to support these accommodations.

Exhibit 6.3.4.1 Accommodations for Students with an Injury

Accommodation	Description
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student’s responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation requires Adult Transcription.
Rest/Breaks	Student may take breaks during testing sessions to rest.

ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNER (ELL) AND FEP STUDENTS

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. Student eligible for these accommodations include English Language Learner (ELL) students,

⁴⁵ Standard 3.10 – When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

⁴⁶ Standard 3.9 – Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees’ ability to demonstrate their standing on the target constructs.

students withdrawn from English language services at parent request and Reclassified Fluent English Proficient (RFEP) students. Students in their monitoring period, within for two school years of reclassifying as Fluent English Proficient (FEP Year 1 and FEP Year 2), may also, as appropriate, use any of the universal test administration conditions and any of the following accommodations.

The accommodations indicated as “*upon student request*” are required to be administered in a setting that does not disturb other students such as in a one-on-one or very small group setting.

Exhibit 6.3.4.2 summarizes accommodations that may be provided for ELL, RFEP, and FEP students.

Exhibit 6.3.4.2 Allowable Accommodations for ELL, RFEP, and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request.</p> <p>Reading aloud the content of the Reading portion of the ELA test.</p>
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	<p>Exact oral translation, in the student’s native language, of the scripted directions or the directions that students read on their own upon student request.</p> <p>Translations that paraphrase, simplify, or clarify directions are.</p> <p>Written translations are.</p> <p>Translation of test content is.</p>
Translation Dictionary	<p>Provide a word-for-word published, paper translation dictionary.</p> <p>Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.</p>

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 6.3.4.3, as designated in their IEP or 504 plan.

Exhibit 6.3.4.3 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
ASL and Closed Caption	CBT – Available for the listening items on the Reading ELA test.
Assistive Technology	Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. This accommodation requires Adult Transcription. Any print copy must be shredded. Any electronic copy must be deleted.
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation also requires Adult Transcription onto a regular size paper test booklet.
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. PBT – Provide a Large Print test booklet. If a Large Print Test Booklet is ordered, an adult must transfer the student's responses exactly as written onto a regular size paper test booklet. This accommodation requires Adult Transcription.
Paper Test Booklet	CBT – Provide a regular size paper test booklet for a student at a school administering the CBT. If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test. Reading aloud the content of the Reading portion of the ELA test.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Sign Test Content	Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the Math test. Signing the content of the Reading portion of the ELA test.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

6.4 SYSTEM SECURITY

6.4.1 SECURE SYSTEM DESIGN

AIR has developed a custom single sign-on application that is made available in Arizona's secure portal. This application is used to support access to AIR's system in accordance with the Arizona's user ID and password policy. Authorized users can log in to Arizona's single sign-on using their current user IDs and passwords and can be redirected to AIR's portal, where they have access to AIR's secure applications such as the Test Information Distribution Engine (TIDE), the test delivery system (TDS), and online reporting system (ORS). Nightly backups protect the data. The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or they will need to rerun the backup. The system can withstand failure of almost any component with little or no interruption of service.

AIR's hosting provider, Rackspace, has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely. Rackspace partners with 9 different network providers, providing multiple, redundant data routes. Every installation is served by multiple servers, any one of which can take over for an individual test upon failure of another.

AIR's architecture ensures data are recoverable at all times. Each disk array is internally redundant, with multiple disks containing each data element. Immediate recovery from failure of any individual disk is performed by accessing the redundant data on another disk. AIR maintains support and maintenance agreements through our hosting provider for all of the hardware used by our systems.

6.4.2 SYSTEM SECURITY COMPONENTS

AIR has built-in security controls in all of its data stores and transmissions.⁴⁷ Unique user identification is a requirement for all systems and interfaces. All of AIR's systems encrypt data at rest and in transit.

⁴⁷ Standard 6.16 – Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information. Standard 8.6 – Test data maintained or transmitted in data files, including all personally identifiable information (not just results), should be adequately protected from improper access, use, or disclosure, including by reasonable physical, technical, and administrative protections as appropriate to the particular data set and its risks, and in compliance with applicable legal requirements. Use of facsimile transmission, computer networks, data banks, or other electronic data-processing or transmittal systems should be restricted to situations in which confidentiality can be reasonably assured. Users should develop and/or follow policies, consistent with any legal requirements, for whether and how test takers may review and correct personal information.

PHYSICAL SECURITY

AzMERIT data resides on servers at Rackspace, AIR's hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. All access is keycard controlled, and sensitive areas require biometric scanning.

Secure data are processed at AIR facilities and are accessed from AIR machines. AIR's servers are in a secure, climate-controlled location with access codes required for entry. Access to our servers is limited to our network engineers, all of whom, like all AIR employees, have undergone rigorous background checks.

Staff at both AIR and Rackspace, receive formal training in security procedures to ensure that they know the procedures and implement them properly. AIR and Rackspace protect data from accidental loss through redundant storage, backup procedures, and secure off-site storage.

NETWORK SECURITY

Hardware firewalls and intrusion detection systems protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) for our secure sites.

AIR's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts.

SOFTWARE SECURITY

All of AIR's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with Arizona's privacy laws, the Family Educational Rights and Privacy Act (FERPA), and other federal laws.

AIR's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. Different states interpret the FERPA differently, and our system is designed to support these interpretations flexibly. AIR has worked with the ADE to maintain data security according to their specifications.

AIR maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results. In addition, AIR runs automated functional tests of our test delivery system every morning, and logs from these runs are available for at least one week from the time of the run.

AIR psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior and also provide a forensics analysis report. The forensics analysis report is described more completely in Section 6.6 on Data Forensics.

6.5 TEST SECURITY

Maintaining a secure test environment is critical to ensure that scores represent what students know and are able to do. Because AzMERIT was administered both as a paper-based and a computer-based assessment, test security procedures must guard against item exposure, cheating on the part of test administrators or students, or other security problems for both testing modes.

The test security procedures involve the following:

- Procedures to ensure security of test materials
- Procedures to investigate test irregularities

Test Administrators are trained on test security procedures and both test security policies and procedures are clearly presented with the *AzMERIT Test Administration Directions*.⁴⁸

Security of Test Materials

All test items, test materials, and student-level testing information are secure documents and must be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment questions, prompts, and student results. Any deviation in test administration must be reported to ensure the validity of the assessment results. Mishandling of test administration puts student information at risk and disadvantages the student. Failure to honor security severely jeopardizes district and state accountability requirements and the accuracy of student data.

The security of all test materials must be maintained before, during, and after test administration. Under no circumstances were students permitted to assist in preparing secure materials before testing or in organizing and returning materials after testing. After any administration, initial or make-up, secure materials (e.g., test booklets, test tickets, used scratch paper) were required to be returned immediately to the School Test Coordinator and placed in locked storage. Secure materials were never to be left unsecured and were not to remain in classrooms or be taken off the school's campus overnight. Secure materials were never to be destroyed (e.g., shredded, thrown in the trash), except for soiled documents. In addition, any monitoring software that would allow test content on student workstations to be viewed or recorded on another computer or device during testing needed to be turned off.

It is unethical and shall be viewed as a violation of test security for any person to:

- capture images of any part of the test via any electronic device;
- duplicate in any way any part of the test;
- examine, read, or review the content of any portion of the test;
- disclose or allow to be disclosed the content of any portion of the test before, during, or after test administration;
- discuss any AzMERIT test item before, during, or after test administration;
- allow students access to any test content prior to testing;

⁴⁸ Standard 6.7 – Test users have the responsibility of protecting the security of test materials at all times.

Standard 7.9 – If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.

- provide any reference sheets to students during the Math test administration;
- allow students to share information during test administration;
- allow students to use scratch paper during the ELA Reading test;
- read any parts of the test to students except as indicated in the Test Administration Directions or as part of an accommodation;
- influence students' responses by making any kind of gestures (for example, pointing to items, holding up fingers to signify item numbers or answer options) while students are taking the test;
- instruct students to go back and reread/redo responses after they have finished their test since this instruction may only be given before the students take the test;
- review students' responses;
- read or review students' scratch paper; or
- participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures.

Additional security violations for paper-based testing include:

- Reading or reviewing any test booklet during or after testing,
- Changing any student response in test booklet,
- Erasing any student's response in test booklet,
- Erasing any stray marks in test booklet,
- Failing to return all test booklets and other test materials.

Test Administrators and Proctors may not assist students in answering questions. Test Administrators and Proctors may not translate, reword, or explain any test content. No test content may ever be discussed before, during, or after test administration.

All regular test booklets and special documents (large print and Braille) test materials are secure documents and must be protected from loss, theft, and reproduction in any medium. A unique identification number and a bar code were printed on the front cover of all test booklets. Schools were expected to maintain test security by using the security numbers to account for all secure test materials before, during, and after test administration until the time they were returned to the contractor.

To access the computer-based AzMERIT tests, a secure Internet browser was required. The secure browser provides a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to the desktop (Internet, email, and other files or programs installed on school machines). The secure browser did not display the IP address or other URL for the site. Users could not access other applications from within the secure browser, even if they knew the keystroke sequences. The "back" and "forward" browser options were not available, except as allowed in the testing environment as testing navigation tools. Students were not able to print from the secure browsers. During testing, the desktop was locked down, and students were required to "Pause" (to save the test for another session) or "Submit" a test in order to exit the secure browser. The secure browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *Test Administrator User Guide* for further details.

Throughout the testing window, test administrators were to report any test incidents (e.g., disruptive students, loss of Internet connectivity) to the School Test Coordinator immediately. A test incident could include testing that was interrupted for an extended period of time due to a local technical malfunction or severe weather. School Test Coordinators notified District Test Coordinators of any test irregularities that were reported. District Test Coordinators were responsible for submitting requests for test invalidations to ADE via AIR's Test Information

Distribution Engine, or TIDE. ADE made the final decision on whether to approve the requested test invalidation. District Test Coordinators could track the status and final decisions of requested test invalidations in TIDE.

6.6 DATA FORENSICS PROGRAM

The validity of test score interpretation depends critically on the integrity of the test administrations on which those scores are based. Any irregularities in the administration of assessments can therefore cast doubt on the validity of the inferences based on those test scores. Multiple facets ensure that tests are administered properly which include clear test administration policies, effective test administrator training, and tools to identify possible irregularities in test administrations.

For online administrations, quality assurance (QA) reports are generated during and after the test windows. These are geared toward detection of testing irregularities that may indicate possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies.

Online test administration allows Arizona's testing contractor to track information that was not possible to track in the context of the paper-and-pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations) as requested by the state. AIR's test delivery system (TDS) captures all of this information.

Unlike with paper assessments where data analysis must await the close of test window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout each test administration window, following the first operational administration. Following the base year, the analyses used to detect the testing anomalies can be run anytime within the testing window. Evidence evaluated included changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

6.6.1 CHANGES IN STUDENT PERFORMANCE

The report examines score changes between years using a regression model. The scores between the previous and current year assessments are compared, with the current-year score regressed on the test score from the previous year.

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. To detect unusual residuals, we compute the studentized t residuals. An unusual increase or decrease in student scores between opportunities is flagged when absolute studentized t residuals are greater than 3.

The number of students with a large score gain or loss is aggregated for a testing session, test administrator, and school. Unusual changes in an aggregate performance between administrations and/or years are flagged based on the average studentized t residuals in an aggregate unit g (e.g., a testing session or a test administrator). For each aggregate unit, a critical t value is computed and flagged when absolute t was greater than 3,

$$t = \frac{\text{Average residuals}}{\sqrt{\frac{s^2}{n_g} + \frac{\sum_{j=1}^{n_g} \text{var}(e_i)}{n_g^2}}}$$

where s = standard deviation of residuals in an aggregate unit; n_g is number of students in the aggregate unit g (e.g., testing session or test administrator); and $\text{var}(e_i) = \sigma^2(1 - h_{ii})$. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

If the aggregate unit size is 1–5 students, the aggregate unit was flagged if the percentage of flagged students was greater than 50%. The aggregate unit size for the score change is based on the number of students included in the within- or between-year regression analyses in the aggregate unit.

6.6.2 ITEM RESPONSE LATENCY

The online environment also allows item response latency to be captured as the item page time (the time each item page is presented) in milliseconds. Discrete items appear one item on the screen at a time. However, for stimulus-based items selected as part of an item group, all items associated with the stimulus are selected and loaded as a group. For each student, the total time taken to complete the test is computed by summing up the page time for all items and item groups.

It is expected that item response time is shorter than the average time if students have prior knowledge of test items. An example of unusual item response time would be a test record for an individual who scores very well on the test even though the average time spent for each item was far less than that required of students statewide. If students already know the answers to the questions, the response time will be much shorter than the response time for those items where the student has no prior knowledge of the item content. Conversely, if a test administrator helps students by “coaching” them to change their responses during the test, the testing time could be longer than expected.

The average and the standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units were flagged if the test-taking time was greater than $|3|$ standard deviations of the state average. The state average and standard deviation was computed based on all students at the time the analysis was performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

6.6.3 INCONSISTENT ITEM RESPONSE PATTERN (PERSON FIT)

In Item Response Theory (IRT) models, person-fit measurement is used to identify examinees whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test-taker has prior knowledge of some test items (or is provided answers during the exam), the student will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response latency index might flag such a student.

The person-fit index is based on all item responses. An unlikely response to a single test question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session and test administrator.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), Sotaridona, Pornell, and Vallejo (2003) define aberrant response patterns as a deviation from the expected item score model. Snijders (2001) showed that the distribution of I_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the “asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05” (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using I_z for systematic flagging of aberrant response patterns. Students with $|I_z|$ values greater than 3 are flagged. Aggregate units are flagged with $|t|$ greater than 3, where t is calculated by

$$t = \frac{\bar{I}_z}{\sqrt{(s^2 + 1)/n_g}},$$

where s is the standard deviation of I_z values in an aggregate unit g and n_g is number of students in the aggregate unit. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit (e.g., test session, test administrator, school).

6.6.4 RESPONSE CHANGE AND RESPONSE SIMILARITY

Response Change in Paper-based Tests

Erasure patterns on paper-pencil tests are also examined for unusual patterns of response changes. For paper-based assessments, we use differences in mark density to infer student erasures, which is then used to identify instances where students may have changed an initial response from incorrect to correct, from incorrect to incorrect, or from correct to incorrect. A set of flagging rules is then used to identify an unusually large number of incorrect to correct erasures at the targeted level of analysis, whether student, testing group, or school. In the online environment, students may change their responses multiple times, and each of those response changes is recorded. Unlike with the mark discrimination analyses, there is no ambiguity about which response was selected or the order in which responses were made. The ease with which response changes can be made, and the accuracy of response capture (i.e., students no longer need to worry that an “erased” response might result in the detection of multiple marks that either cannot be resolved or do not correspond to the student’s intended response) mean that students may now feel freer to change responses, even multiple times for a single item.

Response Pattern Similarity in Computer-based Tests

In fixed-form assessment environments, students may more readily copy from one another than would be possible in a computer adaptive test environment where students are seeing different sets of items in different sequences. To detect possible copying, it can be useful to examine student response records for patterns of excessive response similarity. While similarity in student responses to test questions may be an indicator of irregularities in test administration, response similarity does not always indicate a testing irregularity. For example, in schools with high levels of academic achievement, one would expect large numbers of students to respond correctly, and

therefore similarly, to most items on the test. Nevertheless, patterns of similar responding can indicate testing irregularities, especially when students respond to items incorrectly in the same way. We employ an algorithm, following the model developed by Wesolowsky (2000), for detecting overly similar student responses to multiple-choice items to evaluate patterns of student responses in schools where test irregularities are suspected.

The basic unit of analysis for evaluating response similarity in fixed form assessments is the test session. For each pair of students in a session, we compute the probability of obtaining the same response for each item, including the likelihood of answering the item correctly, as well as selecting the same incorrect response option when answering an item incorrectly. The probability of two students answering an item correctly is conditioned on the average performance of other students in the school. The Bonferroni adjustment is used to correct for the large number of pairwise comparisons, reducing the likelihood of Type I (false positive) errors. A response similarity report identifies pairs of students with overly similar patterns of responding. Exhibit 6.6.4.1 provides sample output for the response similarity analysis. Each record indicates a pair of students flagged for overly similar patterns of responding. Access to a seating chart increases the power of this approach significantly, since students with overly similar response patterns who are known to have been seated in close proximity, obviously have greater opportunity to copy their responses. This method is also useful for detecting cheating rings, where the same students are identified across multiple flagged pairs. This is evident in Exhibit 6.6.4.1, where a common group of students are each flagged in multiple comparisons.

Exhibit 6.6.4.1 Sample Roster Flagging Student Pairs with Excessively Similar Responses

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 LastName	Student1 FirstName	Student2 Barcode	Student2 LastName	Student2 FirstName
SchoolA	Class1	Reading	18		Carter	Adam		Doe	Frank
SchoolA	Class1	Reading	18		Carter	Adam		Farmer	Fred
SchoolA	Class1	Reading	18		Carter	Adam		Miller	Steve
SchoolA	Class1	Reading	18		Carter	Adam		Smith	Cecil
SchoolA	Class1	Reading	18		Carter	Adam		Carter	Henry
SchoolA	Class1	Reading	18		Carter	Adam		Turner	Mark
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Carl
SchoolA	Class1	Reading	18		Carter	Adam		Hall	Robert
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Phillip
SchoolA	Class1	Reading	18		Doe	Frank		Farmer	Fred
SchoolA	Class1	Reading	18		Doe	Frank		Carter	Henry
SchoolA	Class1	Reading	18		Doe	Frank		Hall	Robert
SchoolA	Class1	Reading	18		Doe	Frank		Granger	Phillip
SchoolA	Class1	Reading	18		Farmer	Fred		Miller	Steve
SchoolA	Class1	Reading	18		Farmer	Fred		Smith	Cecil
SchoolA	Class1	Reading	18		Farmer	Fred		Carter	Henry
SchoolA	Class1	Reading	18		Farmer	Fred		Turner	Mark
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Carl
SchoolA	Class1	Reading	18		Farmer	Fred		Hall	Robert
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Phillip
SchoolA	Class1	Reading	18		Miller	Steve		Smith	Cecil
SchoolA	Class1	Reading	18		Miller	Steve		Carter	Henry
SchoolA	Class1	Reading	18		Miller	Steve		Turner	Mark
SchoolA	Class1	Reading	18		Miller	Steve		Hall	Robert
SchoolA	Class1	Reading	18		Miller	Steve		Granger	Phillip

7. REPORTING AND INTERPRETING AZMERIT SCORES

A set of score reports is provided for each administration that summarizes student performance in each grade and content area. Score reports provide data on the performance of individual students and on the aggregated performance of students at various levels — such as state, districts, schools, and teachers. The test data are based on all students who participated in the AzMERIT assessment for the 2015-2016 school year.

The score reports include reliable and valid information describing student progress toward mastery of the state content standards. Arizona provides individual student score reports that are shipped to the student's district for delivery to families. These reports detail student performance on overall tests and subscores. In addition, Arizona offers detailed individual and aggregate level data to educators via AIR's Online Reporting System (ORS), which provides score data for each AzMERIT test, both computer-based and paper-based. The ORS allows users to compare score data between individual students and the school, district, or overall state, and also provides information about performance on subscore categories.

7.1 APPROPRIATE USES FOR SCORES AND REPORTS

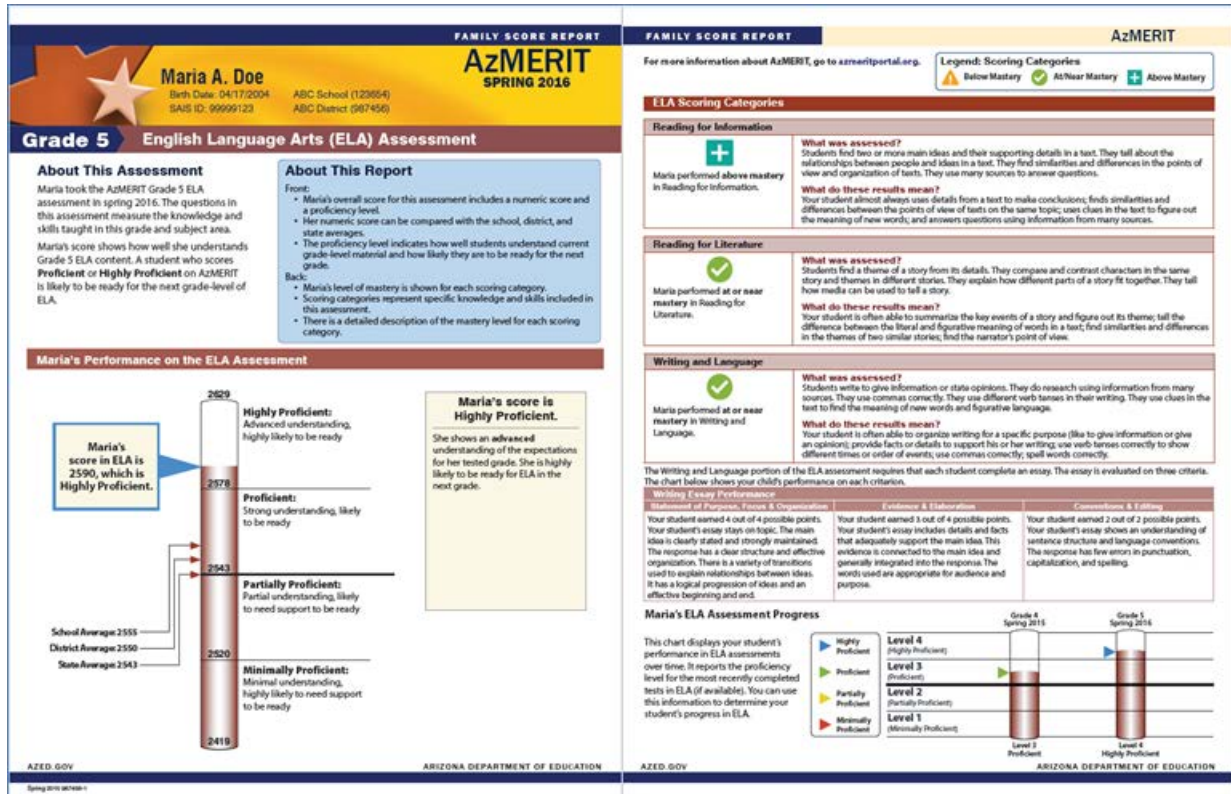
The state provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning and classroom instruction. All reporting systems for the AzMERIT, both paper and online, are designed with stakeholders, such as teachers, parents and students, who are not technical measurement experts, in mind and ensure that test results are used in ways that lead to valid inferences about student achievement and contribute to student learning.⁴⁹ For example, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy guides the reader to compare like elements and avoid comparison of dissimilar elements.

[Sample reports](#) are available at azmeritportal.org. The sections below provide additional guidance for interpreting results.

⁴⁹ Standard 6.10 – When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. Standard 13.5 – Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied.

7.2 REPORTS PROVIDED

7.2.1 FAMILY REPORTS



Arizona provides full-color individual student reports to families of all AzMERIT testers. Reports are designed to be useful to families, and include:

- full color to aid readers' interpretation of the data;
- scale scores and performance level descriptors;
- scoring category performance, including descriptions of what was assessed and what results mean for each scoring category to guide parents and students in their understanding of student scores;
 - A plus (+) symbol indicates that a student is performing above mastery in a particular scoring category
 - A checkmark indicates that a student is performing at or near mastery within the scoring category, and
 - The exclamation symbol indicates a student is performing below mastery in a scoring category
- rubric scores for the writing portion of the ELA test, including descriptions of what those rubric scores mean; and
- school, district, and state average scores for comparative purposes.

In addition, beginning with the spring 2016 administration, ADE provided reports that included longitudinal data as seen at the bottom of the second page of the report. This data is designed to allow parents to track student achievement over time.

7.2.2 ONLINE REPORTING SYSTEM FOR EDUCATORS

AzMERIT results are also reported using AIR's Online Reporting System, which is designed to support educators as they evaluate the needs of their students and reflect on their own curricula and practice. Navigation in the system mirrors the instructional decision-making process, meaning the user can intuitively navigate in any of the three dimensions inherent in the data, helping the user answer three kinds of questions:

1. **Who?** The data can be displayed at levels of aggregation anywhere from the individual level for a specific student up to the entire state. Demographic breakdowns are immediately available at any level of aggregation.
2. **What?** The subject area data can be broken down into finer or coarser "chunks" of content. Navigating this dimension allows the user to travel from subject to scoring category and back.
3. **When?** When data are available over time, the system allows the user to view a data trend over time or toggle to a fixed point in time.

Each navigational step changes the reporting display, providing richer context when interpreting a class's or individual student's performance. While the system contains many reports, the interface design encourages users to think about the substantive, educational questions to which they need answers and access information from that perspective. In addition, while finding and interpreting data from multiple online assessments can easily become overwhelming, the ORS minimizes information overload for educators and administrators by organizing score information in a conceptual framework that helps users quickly locate the right level of data, evaluate its impact, and identify the concrete actions they can take to help students improve.

The AzMERIT online system produces the following online score reports: individual student reports, and aggregate reports at the teacher, school, district, and state level. The AzMERIT online score reports are structured hierarchically. Upon selecting "Home" on the Welcome page, a user is taken to the Home Page Dashboard, which displays for all grades and content areas the number of students tested and the percent of students passing by grade and content area. Users who have access to multiple districts or schools are first required to select a single district or school. Once an aggregate unit is selected in this instance, the summary table of student performance is displayed for the selected entity. For more detailed information for a subject and a grade, the user must select that subject and grade.

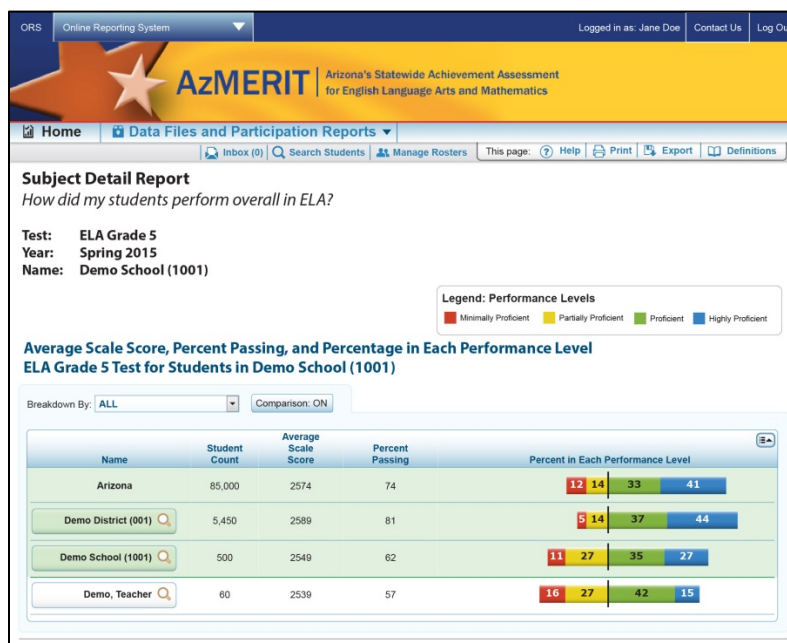
On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district the school belongs to are provided above the school summary results so that the school performance can be compared with the district and the state. If a teacher is selected, the summary results for state, district, and school are provided above the summary results for the teacher.

Exhibit 7.2.2.1 summarizes the types of online score reports available and the levels at which they can be viewed (e.g., student, roster, teacher, school, district).

Exhibit 7.2.2.1 AzMERIT Online Score Report Summary

Type of Report Page	Level of Aggregation	Description
Home Page Dashboard	District, school, and teacher	Summary of performance and participation (Number Tested and Percent Passing) across grades and subjects or course
Subject Detail	District	Average scale score, percent passing, and percent at each performance level for a district and each school within that district; ability to disaggregate data by subgroup
	School	Average scale score, percent passing, and percent at each performance level for a school and each teacher within that school; ability to disaggregate data by subgroup
	Teacher	Average scale score, percent passing, and percent at each performance level for a teacher and each class roster associated with that teacher; ability to disaggregate data by subgroup
Scoring Category Detail	District, school, teacher, and roster	Performance on the scoring category for a subject and a grade for all students and by subgroups; a relative strength and weakness indicator is also reported for each category
Student Roster	School, teacher, roster	List of students with performance on overall subject and scoring categories for a group of students associated with a school, teacher, or roster.
Individual Student Report	Student	Student performance for a selected subject; report includes performance on each scoring category, and performance on the writing essay dimensions, if applicable

SUBJECT DETAIL REPORTS

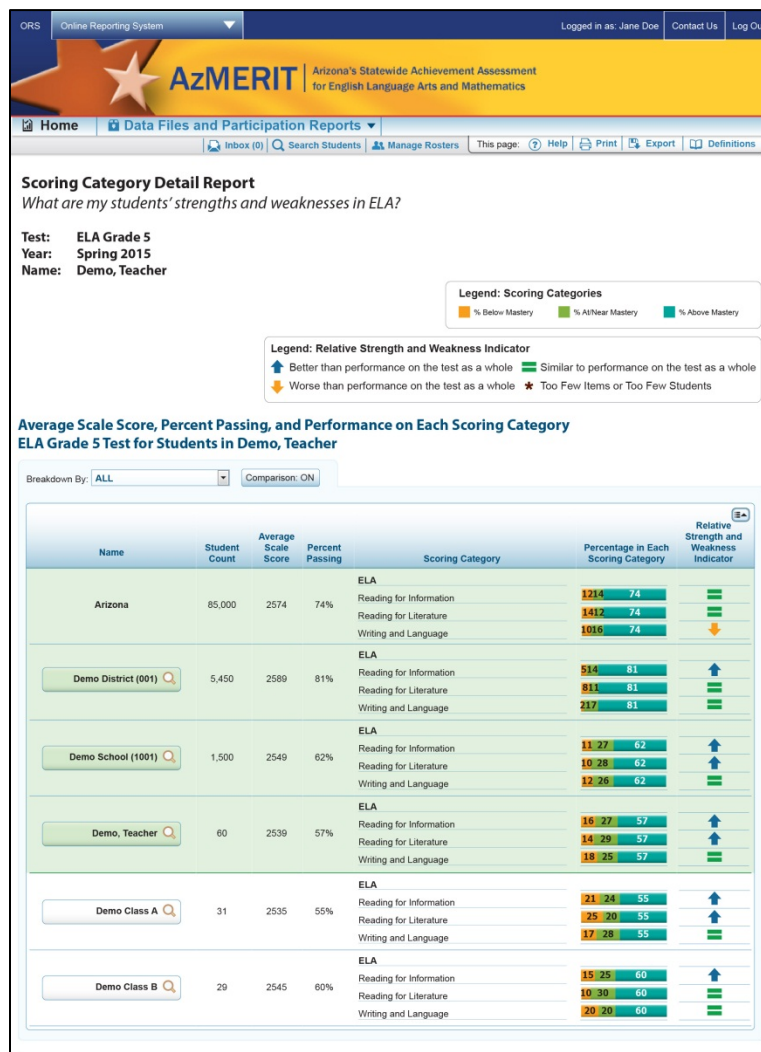


Aggregated subject reports show average performance for the state, districts, schools, teachers, and classes. Bar charts displays show the distribution of students' performance levels. These reports provide users with rosters of schools, teachers, and classes, allowing for simple comparisons across smaller groups.

The Subject Detail Report page shows the following data:

- **Student Count:** Number of students who have completed who completed the selected test
- **Average Scale Score:** Average scale score of students who completed the selected test
- **Percent Passing:** The percent of tested students reaching the proficient threshold on the selected test
- **Percent at Each Performance Level:** The distribution of students across each of the four performance levels

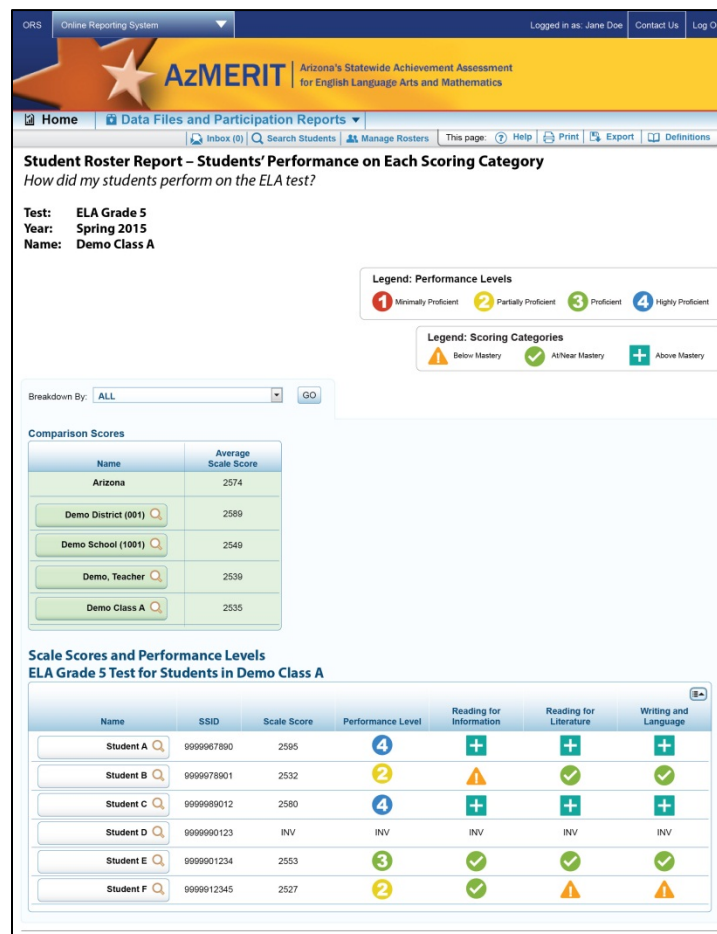
SCORING CATEGORY DETAIL REPORTS



Aggregated scoring category detail reports follow the layout of the subject detail reports, displaying the performance data for the state, districts, schools, teachers, and classes. In addition, these reports include a relative strength and weakness indicator for each category.

In addition to overall test scores, reporting category performance is reported as a strength and weakness indicator. The performance levels indicated on this report are relative to the test as a whole. Unlike performance levels provided at the subject level, these strengths and weaknesses do not imply proficiency. Instead, they show how a group of students' performance is distributed across the scoring categories relative to their overall subject performance on a test. For example, a group of students may have performed very well in a subject, but performed slightly lower in several scoring categories. Thus, the minus sign for a scoring category does not imply a lack of proficiency. Instead, it simply communicates that these students' performance on that scoring category was statistically lower than their performance across all other scoring categories put together. Although the students are doing well, an educator may want to focus instruction on these areas.

STUDENT ROSTER REPORTS



Student roster reports provide users with performance data for a group of students associated with a teacher or a school, as defined in TIDE. The report includes each student's unique state ID, overall subject score, and overall subject performance level. Using the exploration menu, a user can also view each student's scoring category performance for the selected test.

The table that appears on the Student Roster Report page shows the following data:

- **Scale score:** The score of each student who completed the test.

- **Performance level:** Represents levels of overall subject mastery with respect to the Arizona College and Career Ready Standards (4, representing Highly Proficient, to 1, representing Minimally Proficient).
- **Scoring Categories:** Represents levels of scoring category mastery with respect to the Arizona College and Career Ready Standards, characterizing achievement at “above,” “at or near,” or “below” mastery on each scoring category.

INDIVIDUAL STUDENT REPORTS

Individual Student Reports, which closely mirror the Family Reports, are also available through the Online Reporting System.

7.3 INTERPRETATION OF SCORES

Arizona provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning, including interpretive guides for navigating the online reporting system, and understanding paper family reports.⁵⁰ This section describes many of the measures presented in the paper and online score reports.

Performance levels represent levels of mastery with respect to the Arizona College and Career Ready Standards for a content area assessment. Performance levels are labeled as Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance standards are the points on the achievement scale that differentiate performance levels. Three performance standards are used to classify students into one of the four performance levels. Performance standards were recommended by panels of Arizona educators following the first administration of AzMERIT in 2015, and subsequently adopted by the Arizona Board of Education. Panelists engaged in a rigorous, technically sound standard setting process that is summarized in the Performance Standards section of this technical manual, and documented in detail in the 2015 standard setting technical report, available from ADE.

Performance Level Descriptors, or PLDs, define the content area knowledge, skills, and processes that examinees at a performance level are expected to possess. The descriptions of Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient performance are the public statements about what and how much Arizona educators want students to know and be able to do for each grade level and content area. The very detailed PLDs are summarized and included in score reports to provide context for the score and are designed to help parents understand what their students can and cannot do.

The student’s performance in each content area assessment is summarized in an overall test score referred to as a scale score. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale score is then used to determine how well students perform on each content area assessment. Scale scores can be used to measure how much students know and are able to do. Scale scores can also be used to compare student performance across administrations for the same grade and content area so that, for example, an average scale score of 2450 for grade 3 students in the 2015–2016 school year indicates the same level of

⁵⁰ Standard 12.18 – In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

achievement as an average scale score of 2450 for grade 3 students in the 2016–2017 school year even though the test may include a slightly different set of items.

As described in Section 9 on Scaling and Equating, for the ELA assessment, the scale score reported can range from 2395 to 2675. For the math assessment, the scale score reported can range from 3395 to 3839. Overall scale scores for ELA and math are mapped into four performance levels using three performance standards (i.e., cut scores). The AzMERIT scale score ranges can be found in Exhibit 7.3.1.

Exhibit 7.3.1 AzMERIT Scale Score Ranges

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
<i>ELA</i>				
Grade 3 ELA	2395-2496	2497-2508	2509-2540	2541-2605
Grade 4 ELA	2400-2509	2510-2522	2523-2558	2559-2610
Grade 5 ELA	2419-2519	2520-2542	2543-2577	2578-2629
Grade 6 ELA	2431-2531	2532-2552	2553-2596	2597-2641
Grade 7 ELA	2438-2542	2543-2560	2561-2599	2600-2648
Grade 8 ELA	2448-2550	2551-2571	2572-2603	2604-2658
Grade 9 ELA	2454-2554	2555-2576	2577-2605	2606-2664
Grade 10 ELA	2458-2566	2567-2580	2581-2605	2606-2668
Grade 11 ELA	2465-2568	2569-2584	2585-2607	2608-2675
<i>Math</i>				
Grade 3 Math	3395-3494	3495-3530	3531-3572	3573-3605
Grade 4 Math	3435-3529	3530-3561	3562-3605	3606-3645
Grade 5 Math	3478-3562	3563-3594	3595-3634	3635-3688
Grade 6 Math	3512-3601	3602-3628	3629-3662	3663-3722
Grade 7 Math	3529-3628	3629-3651	3652-3679	3680-3739
Grade 8 Math	3566-3649	3650-3672	3673-3704	3705-3776
Algebra I	3577-3660	3661-3680	3681-3719	3720-3787
Geometry	3609-3672	3673-3696	3697-3742	3743-3819
Algebra II	3629-3689	3690-3710	3711-3750	3751-3839

ELA and math assessments are reported on a vertical scale. The item response theory (IRT) vertical scale was developed in 2015 by embedding operational test items from the grade above in the embedded field test slots of each grade level assessment.

8. PERFORMANCE STANDARDS

In the summer of 2015, following the close of the first test administration windows, AIR convened panels of Arizona educators to recommend performance standards on each of the AzMERIT assessments. Details of the panels, procedures, and outcomes are documented in the “Recommending AzMERIT Performance Standards” technical report, which is available from ADE.⁵¹ This section briefly describes the procedures used by educators to recommend standards, and resulting performance standards.

8.1 STANDARD SETTING PROCEDURES

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the Arizona College and Career Ready Standards. AzMERIT test scores are reported with respect to the four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the Arizona College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Arizona College and Career Ready Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standard for the AzMERIT assessments are therefore central to the validity of test score interpretations.

Following the first operational administration of the AzMERIT assessments in spring 2015, a standard setting workshop was conducted to recommend to the Arizona State Board of Education a set of performance standards for reporting student achievement of the Arizona College and Career Ready Standards. The workshop consisted of a series of standardized and rigorous procedures that the Arizona educators serving as standard setting panelists followed to recommend performance standards. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists use their expert knowledge of the Arizona College and Career Ready Standards and student achievement to map the performance level descriptors adopted by the Arizona State Board of Education onto an ordered item book (OIB) based on the first operational test form administered to students in spring 2015.

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were

⁵¹ Standard 5.21 – When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Standard 7.4 – Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.

provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

Panelists were also provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade level assessment related to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces the interpretation of test scores as indicating not only achievement of current grade level standards, but also preparedness to benefit from instruction in the subsequent grade level.

8.1.1 PERFORMANCE LEVEL DESCRIPTORS

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance level descriptors (PLDs) define the content area knowledge and skills that students at each performance level are expected to demonstrate. The standard setting panelists based their judgments about the location of the performance standards on the PLDs as well as the Arizona College and Career Readiness Standards. The AzMERIT PLDS describe four levels of achievement:

- Minimally Proficient
- Partially Proficient
- Proficient
- Highly Proficient

Prior to convening the standard setting workshops, AIR, in consultation with ADE, drafted PLDs for each test that described the range of achievement encompassed by each performance level on the test. The PLDs were designed to be clear, concrete, and reflect Arizona's expectations for proficiency based on the Arizona College and Career Ready Standards. Following a cycle of revisions to the draft PLDs, ADE invited Arizona educators to review PLDs for each of the assessments. Based on feedback from 166 educators, PLDs were further revised, and the resulting drafts were used by standard setting panelists. ADE considered any need for clarification or revision that arose throughout the standard setting process prior to publishing the final versions of the PLDs following the standard setting workshop. [AzMERIT PLDs](#) are available at azed.gov.

8.2 RECOMMENDED PERFORMANCE STANDARDS

Panelists were tasked with recommending three performance standards (Partially Proficient, Proficient, and Highly Proficient) that resulted in four performance levels (Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient). Exhibit 8.2.1 presents the performance standard associated with panelist-recommended OIB page numbers in logit value, as well as the percentage of students classified as meeting or exceeding each standard. Following the standard setting workshop, panelist recommendations were submitted to Arizona's State Board of Education; the Board formally adopted the standards in August 2015.

Exhibit 8.2.1 Final Recommended Performance Standards for AzMERIT

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% At or Above	Theta	% At or Above	Theta	% At or Above
ELA						
3	-0.09	56	0.29	41	1.36	10
4	0.14	57	0.6	39	1.8	5
5	-0.13	63	0.63	30	1.8	3
6	-0.12	61	0.58	34	2.03	4
7	-0.02	59	0.61	33	1.9	4
8	-0.06	60	0.64	33	1.72	6
9	-0.12	53	0.59	27	1.57	6
10	0.11	51	0.58	30	1.42	8
11	-0.02	46	0.52	26	1.27	8
Math						
3	-0.16	73	1.04	42	2.43	15
4	-0.31	71	0.76	42	2.2	10
5	-0.65	71	0.41	40	1.74	13
6	-0.48	62	0.41	32	1.55	11
7	-0.19	52	0.59	30	1.51	13
8	-0.69	57	0.09	32	1.15	13
Algebra I	-0.69	55	-0.03	32	1.27	9
Geometry	-1.37	53	-0.58	30	0.96	6
Algebra II	-1.49	53	-0.78	29	0.57	6

Exhibit 8.2.2 shows the percentage of student classified at each performance level in the initial year of AzMERIT administration, based on final panelist-recommended standards for the student population overall across grade levels and courses for the ELA and math assessments.

Exhibit 8.2.2 Percentage of Students at Each Performance Level based on Final Recommended Performance Standards

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
ELA				
3	44%	15%	31%	10%
4	43%	19%	33%	5%
5	37%	33%	27%	3%
6	39%	27%	30%	4%
7	41%	26%	29%	4%
8	40%	27%	26%	6%
9	47%	26%	21%	6%
10	49%	21%	22%	8%
11	54%	20%	17%	8%
Math				
3	27%	31%	27%	15%
4	29%	29%	32%	10%
5	29%	31%	27%	13%
6	38%	30%	21%	11%
7	48%	22%	18%	13%
8	43%	24%	20%	13%
Algebra I	45%	23%	23%	9%
Geometry	47%	24%	24%	6%
Algebra II	47%	24%	23%	6%

Exhibit 8.2.3 shows the percentage of students meeting the AzMERIT proficient standard for each assessment in the base year of 2015 (meaning they are categorized as Proficient or Highly Proficient), and the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8, and the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2015 field test administration. As Exhibit 8.2.3 indicates, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college ready, and the NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 8.2.3 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Grade/ Course	Percent of Students Meeting Standard			
	AzMERIT Proficient	Arizona ACT College Ready	Arizona NAEP Proficient	Projected SBAC
<i>ELA</i>				
3	41%			38%
4	38%		28%	41%
5	30%			44%
6	34%			41%
7	33%			38%
8	32%		28%	41%
9	27%			
10	30%			
11	25%	34%		41%
<i>Math</i>				
3	42%			39%
4	42%		42%	38%
5	40%			33%
6	32%			33%
7	31%			33%
8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

9. SCALING AND EQUATING

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^n P(z|\theta),$$

where Z represents the pattern of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function $P(Z)$. The one-parameter model (1PL; also known as the Rasch model), is used to calibrate AzMERIT items that are scored either right or wrong, and takes the form

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i .

The b parameter is often called the *location* or *difficulty* parameter; the greater the value of b , the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i|\theta) = \frac{\exp \sum_{k=0}^{x_i} (\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l (\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^0 (\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i . Item parameters for the assessments were calibrated following the spring administration in 2015 and vertical scales were established for reporting both ELA and math. In addition, a series of linking studies were performed to allow the comparison of performance on the AzMERIT to other state and national scales. A mode comparability study was also completed to examine possible effects of test administration mode. These studies were completed prior to establishing performance standards in summer 2015 and subsequent scoring and reporting of AzMERIT results. AzMERIT ELA is reported on a scale ranging from 2395 to 2675 across the grade level and high school End-of-Course tests. AzMERIT math is reported on a scale ranging from 3395 to 3839 across grade level and high school End-of-Course (Algebra I, Geometry, and Algebra II) tests.

9.1 ITEM RESPONSE THEORY PROCEDURES

The AzMERIT assessment was administered for the first time in the spring of 2015. Following test administration, item response theory (IRT) procedures were used to calibrate item parameter estimates and create the new AzMERIT scales for scoring and reporting.⁵² This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by AIR, ADE, and HumRRO, which acts as a third party quality assurance contractor.

Within AzMERIT, students are able to skip items in both the online and paper test platforms. While omitted items are scored as incorrect for purposes of ability estimation, all omitted responses are treated as not-administered for purposes of IRT analysis. All students who respond to at least one item within each test session are considered to have attempted a test. All attempted records are included in IRT analysis with the exclusion of students who had more than one record for the same test and records that are had been invalidated prior to scaling.

9.1.1 CALIBRATION OF AZMERIT ITEM BANKS

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for AzMERIT. Winsteps is publically available software from Mesa Press. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters' (1982) partial credit model, an extension of the one parameter Rasch model which allows for partial credit to be given on items, estimates the responses for polytomous items.

In spring 2015, operational items for each test were freely calibrated establishing the new AzMERIT reference scales. Following the approval of final item parameter estimates for operational items, parameter estimates for the operational items were anchored to their new AzMERIT bank values and parameter estimates for field test and linking items were estimated under that constraint. This placed parameter estimates for all field test and external linking items on the same AzMERIT scale defined by the operational item parameters.

In spring 2016, pre-equated item parameters were used to score student test records for the math assessments. For ELA, since six new writing tasks at each grade were being administered in the ELA assessments, operational ELA items were recalibrated, and the equating constant necessary to place the common items back to the reference scale was identified and applied to the recalibrated item parameters. This placed all test items on the base year AzMERIT scale. Mean equating was used to compute the linking constant, and all operational reading items were included in the linking computation.

⁵² Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

9.1.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

To identify the likelihood of a student's ability across the ability distribution, we begin by evaluating the likelihood of achieving a score point for an item given the underlying level of ability. Let X_i be a random variable taking a student's response on item i ($i = 1, \dots, N$) with an outcome $x_i \in \{0, 1, \dots, m_i\}$. Item i is a dichotomously scored item if $m_i = 1$, and polytomously scored item if $m_i > 1$. Based on Masters' (1982) partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i | \theta) = \frac{\exp \sum_{k=0}^{x_i} (\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l (\theta - b_{kl})},$$

with the constraint that $\sum_{k=0}^0 (\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i . Note that if item i is a dichotomously scored item, the partial credit model becomes the Rasch model and can be written as

$$P(X_i = 1 | \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i .

LIKELIHOOD FUNCTION

The likelihood function of ability θ given responses to N items, $\mathbf{x} = \{x_i\}$, can be expressed as:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^N P(x_i | \theta).$$

The maximum likelihood estimate $\hat{\theta} = \arg \max_{\theta} L(\theta | \mathbf{x})$ or equivalently, $\hat{\theta} = \arg \max_{\theta} \ln L(\theta | \mathbf{x})$.

DERIVATIVES

Finding the maximum likelihood estimate requires an iterative method, such as Newton-Raphson iterations. Since the log-likelihood is a monotonic function of the likelihood, the following derivatives based on the log-likelihood function are used:

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \theta} &= \sum_{i=1}^N \left[x_i - \sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right] \\ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} &= \sum_{i=1}^N \left[\sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right]^2 - \sum_{i=1}^N \sum_{x_i=0}^{m_i} x_i^2 P(X_i = x_i | \theta) \end{aligned}$$

The maximum likelihood estimates of θ is found via the following iterative routine:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\partial \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t} / \frac{\partial^2 \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t^2}.$$

This iterative process repeats until the difference between $\hat{\theta}_t$ and $\hat{\theta}_{t+1}$ is less than a pre-specified threshold.

ESTIMATING ZERO AND PERFECT SCORES

In the event of zero or perfect scores, a procedure recommended by Berkson (as cited in Linacre, 2004) is implemented to add (or subtract) 0.5 to (or from) the test score prior to estimating student ability. Thus, for students responding incorrectly to all items in a scale or subscale, students will be assigned a test score of 0.5. Conversely, for students responding correctly to all items in a scale or subscale, 0.5 will be subtracted from the raw score prior to calibration.

9.2 ESTABLISHING A VERTICAL SCALE IN ELA AND MATH

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade level assessment (g) were embedded in field test slots of the assessment in the grade below ($g - 1$).⁵³ In this approach, the resulting linkage represents student achievement each year on the scale of the subsequent grade level assessment for which they are preparing to receive instruction. As such, the scale scores for each assessment can be interpreted as a pre-test score for measuring student acquisition of academic content in the subsequent grade level. While this approach risks administering to students 1-2 items measuring content that they may not yet have had the opportunity to learn, it provides a more sensitive measure of student growth than could be obtained by a linking design in the linkage represents continued growth on academic content assessed in the previous year's assessment.

9.2.1 LINKING ITEMS

Since the vertical scale essentially places each AzMERIT assessment on the scale for the assessment in the grade above, we can best assure comparability of test scores between the grades by establishing the linkage using all available operational test items. Thus, to link the grade 4 assessments to the grade 5 scales, all operational items in the grade 5 assessment were made available for administration in the grade 4 embedded field test (EFT) slots. The inclusion of all operational items in the vertical linking set ensures that the item set used to link to the target adjacent grade scale fully represents the measured construct in the target grade, allowing for valid inferences to be made with respect to student baseline performance for achievement in the subsequent grade level.

Because the AzMERIT assessments of English language arts (ELA) in high school continue as End-of-Course (EOC) or grade-level measures of student achievement of the Arizona College and Career Ready Standards (ACCRS), each assessment can be linked to the grade above using all available operational items.

However, AzMERIT assessments of high school math are composed of a set of EOC tests that are not as consistently associated with grade-level instruction and which measure specific subsets of the content domain. For example, while math coursework in high school follows a typical progression and it would therefore be possible to embed “grade 9” Algebra I EOC items in the grade 8 math assessment, embed the “grade 10” Geometry EOC items in the Algebra I EOC exam, and embed the “grade 11” Algebra II the Geometry exam, the constructs measured

⁵³ Standard 5.0 – Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.

Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

across the four exams vary considerably and have implications for the interpretation of growth, or lack thereof, across assessments. For example, it is not clear what the expectation for growth should be in a vertical scale established by embedding Geometry items in an Algebra I exam, since Geometry is not a focus of instruction in Algebra I courses. An alternative approach, and the one adopted by ADE, was to link the grade 8 math scale to both the Algebra I and Geometry EOC scales, since the grade 8 assessment includes items measuring both algebra and geometry. Because Algebra II builds on the knowledge and skills assessed in Algebra I, all Algebra II items were used to link the Algebra I assessments to the Algebra II scale.

9.2.2 LINKING ANALYSIS

When feasible, it is desirable to establish linkages using both concurrent calibrations and chain-linking approaches to ensure that results are consistent across methods. An important advantage of chain linking approaches is that, because item response theory (IRT) calibrations proceed by establishing the within-grade scale, the achievement construct intended by the blueprint and enacted in the operational test form is preserved. Unfortunately, however, at each step in the linking chain, the linking error accumulates, so that linking constants for grades more distant from the reference grade are less precise than are linking constants for grades in closer proximity to the reference grade. Concurrent calibrations do not accrue linking error across grade levels, so that linking constants are similarly precise between all grade levels. However, the calibrations resulting from this approach measure the construct that is common across the linked assessments, which may be different from the intended achievement construct at each grade level, especially for subjects such as math where the assessed construct may change markedly across grade levels. Generally, both approaches tend to converge to produce vertical scales that operate similarly (Ito, Sykes, and Yao, 2008; Karkee, Lewis, Hoskens, Yao, and Haug, 2003), and we view convergence as evidence for the robustness of the vertical scale.

Final Linking Set

To facilitate the development of a vertical scale that will be sensitive to student growth over time, we first evaluated the performance of vertical linking items between the grade levels in which they were administered to identify any items that were more difficult for students in the intended grade than they were for students in the lower grade. For math, items that showed proportion correct scores lower in the intended grade than in the lower grade were dropped from the final vertical linking set. This resulted in dropping on average just over two items per linking set, with a maximum of six items dropped for the linkage between grade 6 and grade 7 math assessments.

For reading, the proportion correct values across grades were much closer, especially at the higher grade levels, so that elimination of all items where the proportion correct value in the lower grade exceeded the higher grade would result in dropping more items from the vertical linking set than would be desirable for executing a robust equating design. Thus, we modified the rule for reading to exclude from the vertical linking set those items which showed proportion correct values more than two standard errors beyond the average standard error for the total linking set (i.e., items that were reliably less difficult at the lower grade). This approach allowed us to identify a final set of linking items that would maximize detection of growth, while retaining sufficient items to establish a strong linkage between the grade level assessments.

Exhibit 9.2.2.1 Number of Items Dropped and Remaining in the Final Vertical Linking Set

Linkage	Math Dropped Items	Math Final VL Set	ELA Dropped Items	ELA Final VL Set
G3→G4	1	44	1	42

Linkage	Math	Math	ELA	ELA
	Dropped Items	Final VL Set	Dropped Items	Final VL Set
G4→G5	0	45	3	46
G5→G6	1	46	0	47
G6→G7	6	41	5	39
G7→G8	3	47	2	46
G8 M→ Algebra I & G8 ELA→G9 ELA	3	28	11	30
G8 M→Geometry & G9 ELA→ G10 ELA	2	31	7	39
Algebra I→ Algebra II & G10 ELA→ G11 ELA	2	32	10	35

CHAIN-LINKING

The chain linking approach proceeds from the within grade item parameters identified in the initial calibrations of the operational and embedded field test items. Because operational test items at each grade were administered in the EFT slots in the grade below, each item in the vertical linking set has two sets of item parameters: on-grade (g) and below-grade (g - 1). The chain linking proceeds by identifying the linking constants necessary to place the below-grade item parameters onto the on-grade scale for the items in the final vertical linking set. The linking constant for each grade was defined as the mean difference of the item difficulty estimates for the linking items between the linked grades. The chain linking began by placing the grade 3 item parameters on the grade 4 scale for both math and ELA and proceeded upwards. For math EOC assessments, the grade 8 math scale was linked to both the Algebra I and Geometry scales, and the Algebra I scale was linked to the Algebra II scale.

CONCURRENT CALIBRATION

A vertical scale for each subject area was also established by calibrating simultaneously all items in the final vertical linking set. As with the within grade calibrations, parameters were estimated using Winsteps. To compare results from the chain-linking and concurrent calibrations, the concurrent calibrations were placed on the grade 3 reference scale.

Exhibit 9.2.2.2 shows the vertical linking constants resulting from chain-linking the within grade scales as well as from concurrently calibrating items from across grade levels. The linking constants are applied to their respective within grade scale to place all item parameters on the grade 3 reference scale. To more directly examine the magnitude of gains across grade level assessments, Exhibit 9.2.2.3 shows the difference between linking constants between each of the grade levels assessed. Relative gains are also represented graphically in Exhibit 9.2.2.4 and Exhibit 9.2.2.5 for math and ELA, respectively, which plot the linking constants across grade level assessments. As the linking constants indicate, for math there is relatively large and steady growth across the grade level and end of course assessments. For the ELA assessments, the cross grade gains are more modest, and tend to diminish in the higher grade levels.

Exhibit 9.2.2.2 Vertical Linking Constants Resulting from Chain-Linking Within Grade Scales and Concurrent Calibration of Items Across Grades

Linkage	Math	Math	ELA	ELA
	Chain Linked	Concurrent	Chain-Linked	Concurrent
G3→G4	1.32	1.30	0.18	0.16
G4→G5	2.75	2.67	0.81	0.78
G5→G6	3.90	3.73	1.19	1.15
G6→G7	4.48	4.28	1.44	1.39

Linkage	Math		ELA	
	Chain Linked	Concurrent	Chain-Linked	Concurrent
G7→G8	5.69	5.39	1.76	1.70
G8 M→ Algebra I & G8 ELA→G9 ELA	6.07	5.76	1.97	1.88
G8 M →Geometry & G9 ELA→ G10 ELA	7.15	6.86	2.12	1.98
Algebra I→ Algebra II & G10 ELA→ G11 ELA	7.81	7.45	2.32	2.16

Exhibit 9.2.2.3 Linking Constant Differences Between Each of the Grade Level Scales

Linkage	Math Chain Linked	Math Concurrent	ELA Chain-Linked	ELA Concurrent
G3→G4	1.32	1.30	0.18	0.16
G4→G5	1.43	1.37	0.63	0.62
G5→G6	1.15	1.06	0.38	0.37
G6→G7	0.58	0.55	0.25	0.24
G7→G8	1.21	1.11	0.32	0.31
G8 M→ Algebra I & G8 ELA→G9 ELA	0.38	0.37	0.21	0.18
G8 M →Geometry & G9 ELA→ G10 ELA	1.08	1.10	0.15	0.10
Algebra I→ Algebra II & G10 ELA→ G11 ELA	0.66	0.59	0.20	0.18

Linking constants resulting from the chain-linking and concurrent calibration approach are quite consistent, indicating that both approaches converge on a common growth scale. Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain-linking method preserves the within grade measurement construct, and was therefore selected as a preliminary vertical scale for the purpose of recommending performance standards. We note that ordered item books for the standard setting workshop were based on the within grade scales, so any modifications to the vertical scale would not impact the recommended performance standards.

The vertical linking constants also indicate much greater growth across grades and high school courses for math than is observed for ELA. In math, growth is on the order of about one standard deviation per year, with the exception of grade 6 to grade 7, which showed just over a half standard deviation gain. Similar half standard deviation gains were observed between grade 8 and Algebra I, which some students take concurrently, and between coursework in Algebra I and Algebra II. Gains in ELA are less pronounced, with somewhat larger gains in the elementary school years, with growth attenuating in the high school grades.

Exhibit 9.2.2.4 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: Math

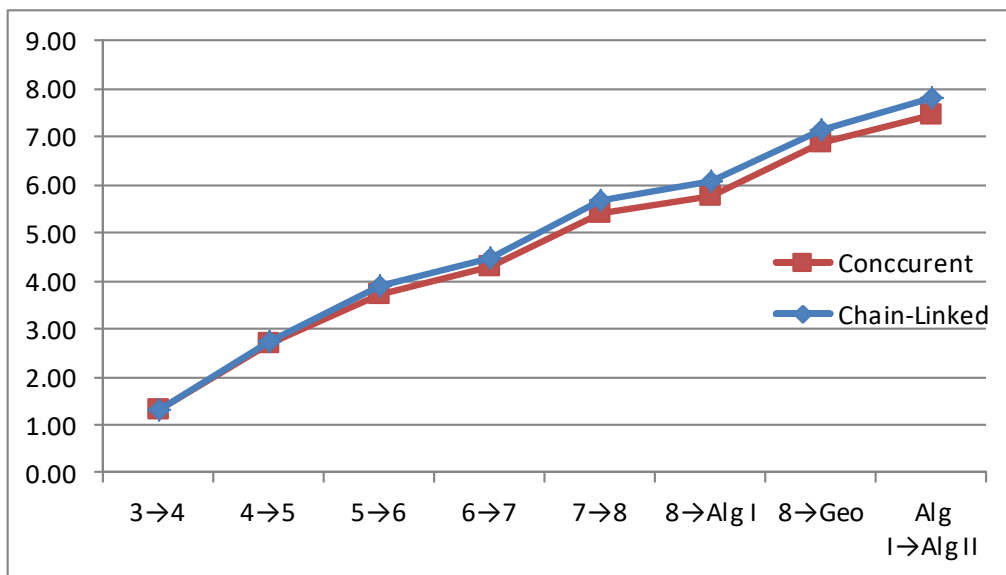
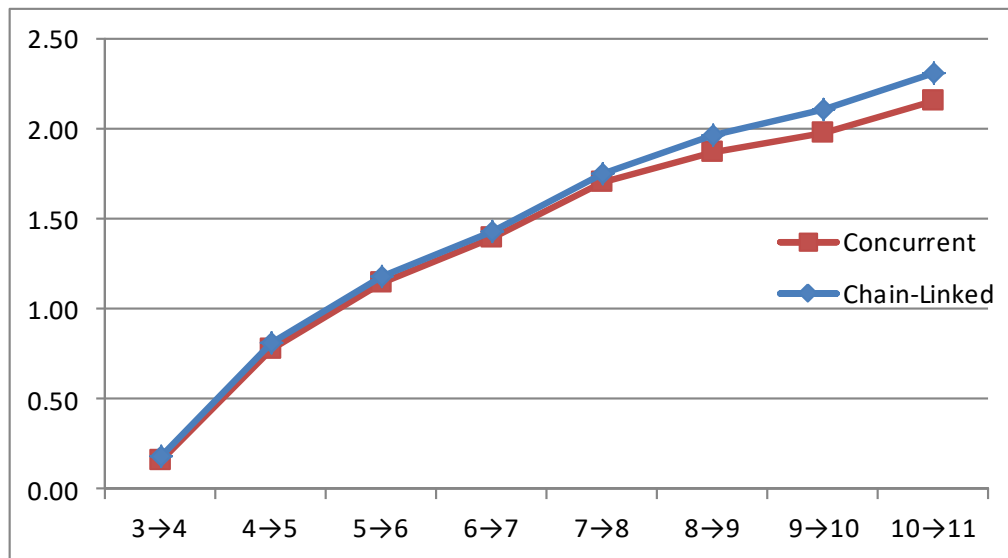


Exhibit 9.2.2.5 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: ELA



9.3 AZMERIT REPORTING SCALE (SCALE SCORES)

The AzMERIT assessments are reported on common scales within each subject (ELA and math). The IRT vertical scale scores (SS) are formed by linking each grade level assessment to the scale of the assessment in the grade level above. The vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate,⁵⁴

$$SS = a * \theta_v + d$$

where $a = 30, d = 2500$ for ELA tests, and $a = 30, d = 3500$ for Math tests. $\theta_v = \theta + c$, where θ is the on-grade ability estimate and c is a vertical linking constant listed below for each of the tests, as described in the previous section. For reporting, the on-grade ability estimate is truncated at ± 3.5 .

After transforming theta ability estimates to the vertical AzMERIT reporting scale, the observable scale scores nearest each of the performance standard cut scores are evaluated. If the observable scale score nearest the performance standard is below the cut score, the scale score is rounded up to be equal to the cut score. If the observable scale score nearest the performance standard is above the cut score, no special rounding rule is applied.

Overall scale scores for the AzMERIT are mapped into 4 performance levels per grade/course. The performance level designations are: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. The performance level is evaluated using the rounded scale score.

Exhibit 9.3.1 shows the scale score ranges for the performance levels for each test.

⁵⁴ Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Exhibit 9.3.1 Scale Score Ranges for Performance Levels

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
<i>ELA</i>				
Grade 3 ELA	2395-2496	2497-2508	2509-2540	2541-2605
Grade 4 ELA	2400-2509	2510-2522	2523-2558	2559-2610
Grade 5 ELA	2419-2519	2520-2542	2543-2577	2578-2629
Grade 6 ELA	2431-2531	2532-2552	2553-2596	2597-2641
Grade 7 ELA	2438-2542	2543-2560	2561-2599	2600-2648
Grade 8 ELA	2448-2550	2551-2571	2572-2603	2604-2658
Grade 9 ELA	2454-2554	2555-2576	2577-2605	2606-2664
Grade 10 ELA	2458-2566	2567-2580	2581-2605	2606-2668
Grade 11 ELA	2465-2568	2569-2584	2585-2607	2608-2675
<i>Math</i>				
Grade 3 Math	3395-3494	3495-3530	3531-3572	3573-3605
Grade 4 Math	3435-3529	3530-3561	3562-3605	3606-3645
Grade 5 Math	3478-3562	3563-3594	3595-3634	3635-3688
Grade 6 Math	3512-3601	3602-3628	3629-3662	3663-3722
Grade 7 Math	3529-3628	3629-3651	3652-3679	3680-3739
Grade 8 Math	3566-3649	3650-3672	3673-3704	3705-3776
Algebra I	3577-3660	3661-3680	3681-3719	3720-3787
Geometry	3609-3672	3673-3696	3697-3742	3743-3819
Algebra II	3629-3689	3690-3710	3711-3750	3751-3839

9.4 LINKING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to reporting test scores for the spring 2015 and spring 2016 administrations of AzMERIT, AIR and ADE performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration.⁵⁵

9.4.1 MODE LINKING

A matched samples design (Way, Davis, and Fitzpatrick, 2006) was used to investigate mode comparability. A covariate regression approach was implemented to construct equivalent groups of students taking the AzMERIT assessments for both modes of test administration. For the spring 2015 mode investigation, the regression analysis identified for each student a predicted score on the paper AzMERIT assessment from previous year achievement on AIMS, covarying demographic variables that included gender, ethnicity, income level status, English language learner (ELL) status, and Individualized Education Program (IEP) in the development of the prediction equation. A

⁵⁵ Standard 5.13 – When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

nearest neighbor search procedure was then applied to the predicted AzMERIT scores to select the equivalent groups of students. This procedure resulted in the identification of two matched samples for each assessment to conduct the mode comparability study.

IRT parameter estimates were then calibrated independently for the matched online and paper test administration mode samples. The linking constant necessary to bring the matched sample paper item parameters onto the matched sample online scale was then computed. Mean-mean linking was taken as the difference between the average item difficulty estimates from the matched-sample paper calibration and the average item difficulty estimates from the matched-sample online item parameter estimates.

Mode linking constants were estimated again following the spring 2016 administration of AzMERIT. Three approaches were used to identify matched samples for these analyses. In the first approach, 2014 AIMS paper test scores were used to predict student performance on the spring 2016 paper tests, with the resulting prediction model then used to identify a matched sample of online test takers. This approach allowed all available paper records to be included in the analysis, but required constructing matched samples based on achievement scores estimated two years prior. To utilize a more recent and comparable test score, a second approach was used. In this approach, we identified students who were administered AzMERIT on paper in 2015, but who participated online in spring 2016. We then identified a matched sample of students, based on AzMERIT test scores, who took the paper version of AzMERIT in both 2015 and 2016. For students at grade 3, there were no previous test scores with which to match student ability. We therefore used student performance on the multiple-choice items only on the spring 2016 AzMERIT math test to identify matched samples on the assumption that those items would be least susceptible to mode differences. To evaluate whether this approach yields results consistent with the other approaches, this approach was also applied to the grade 4 and grade 5 assessments.

Exhibit 9.4.1 presents the mode linking constants for the ELA assessments resulting from the matched sample analysis conducted on the spring 2015 administration of AzMERIT, as well as the linking constants resulting from each of the matched sample approaches used following the spring 2016 administration. In the grade 4 through 8 assessments, whether the matched samples are based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode linking constants are generally small and equivalent across methods. For the high school end-of-course assessments, both approaches indicate that ELA assessments were somewhat more difficult online than on paper. The magnitude of those differences is greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. We note that the R^2 for the prediction equation used to identify matched samples for ELA based on 2014 AIMS remained quite high (R^2 around 0.65) even for the high school assessments, although matching based on spring 2015 AzMERIT achievement may nevertheless be more robust.

For grade 3 ELA, samples were matched based on student performance on the concurrently administered AzMERIT math multiple-choice items. To evaluate whether this approach yielded results consistent with the other two methods, we applied the same procedure in grades 4 and 5, where results indicated general convergence with the other methods, and indicating no effect for mode at grade 4 and a moderate mode effect at grade 5. When applied at grade 3, no effect for mode was identified.

We note that any effect of mode seems to interact with items, with some items easier when administered online, while others are more difficult. Thus, the effect of mode is likely to be form specific and vary across test administrations. And this seems to be the case when mode linking constants are compared between the 2015 and 2016 administrations of AzMERIT. As shown in Exhibit 9.4.1, in spring 2015, mode effects were observed in grades 3, 4, and 8, but were more moderate at the other grades. In spring 2016, however, mode effects were absent or moderate in grades 3 through 8, but appear in the high school End-of-Course tests.

Exhibit 9.4.1 Mode Linking Constants for AzMERIT ELA Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3E	2015	0.13	-0.01	0.13	3.90
	2016 – Math MC Match	0.17	0.16	0.01	0.30
G4E	2015	-0.09	-0.19	0.11	3.30
	2016 – 2014 AIMS Match	0.21	0.19	0.02	0.60
	2016 – 2015 AzMERIT Match	0.21	0.18	0.03	0.90
	2016 – Math MC Match	0.21	0.21	0.00	0.00
G5E	2015	0.04	-0.02	0.06	1.80
	2016 – 2014 AIMS Match	0.02	-0.02	0.04	1.20
	2016 – 2015 AzMERIT Match	0.03	-0.02	0.05	1.50
	2016 – Math MC Match	0.04	-0.04	0.08	2.40
G6E	2015	0.07	-0.02	0.09	2.70
	2016 – 2014 AIMS Match	0.18	0.21	-0.03	-0.90
	2016 – 2015 AzMERIT Match	0.20	0.16	0.04	1.20
G7E	2015	-0.08	-0.16	0.08	2.40
	2016 – 2014 AIMS Match	0.19	0.12	0.07	2.10
	2016 – 2015 AzMERIT Match	0.12	0.05	0.07	2.10
G8E	2015	-0.04	-0.22	0.18	5.40
	2016 – 2014 AIMS Match	0.01	-0.01	0.02	0.60
	2016 – 2015 AzMERIT Match	0.00	-0.05	0.05	1.50
G9E	2015	0.13	0.09	0.04	1.20
	2016 – 2014 AIMS Match	0.07	-0.12	0.20	6.00
	2016 – 2015 AzMERIT Match	0.08	-0.16	0.24	7.20
G10E	2015	-0.03	-0.10	0.07	2.10
	2016 – 2014 AIMS Match	0.10	-0.10	0.20	6.00
	2016 – 2015 AzMERIT Match	0.09	-0.04	0.13	3.90
G11E	2015	0.12	0.15	-0.03	-0.90
	2016 – 2014 AIMS Match	0.16	-0.09	0.25	7.50
	2016 – 2015 AzMERIT Match	0.14	-0.04	0.18	5.40

Exhibit 9.4.2 presents the mode linking constants computed for the spring 2015 and spring 2016 administrations of the AzMERIT math assessments. As observed for ELA, in the grade 4 through 8, and Algebra I math assessments, whether the spring 2016 matched samples were based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode linking constants are generally equivalent across methods. Effects of mode varied across grades, with the online form somewhat easier than paper at grade 4, somewhat more difficult at grade 7, and about the same at grades 5, 6, and 8. For the high school end-of-course assessments, both approaches indicate that math assessments were somewhat more difficult online than on paper. As with ELA, the magnitude of those differences was greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. In this case we note that the R^2 for the prediction equation used to identify matched samples for math based on 2014 AIMS remained quite a bit

lower ($R^2 \approx .40$) for the high school assessments compared to the lower grades ($R^2 \approx .65$), so that matching based on spring 2015 AzMERIT achievement are likely more robust.

Exhibit 9.4.2 Mode Linking Constants for AzMERIT Math Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3M	2015	-0.71	-0.77	0.06	1.80
	2016 – Math MC Match	-0.84	-0.57	-0.27	-8.10
G4M	2015	-0.40	-0.48	0.08	2.40
	2016 – 2014 AIMS Match	-0.43	-0.25	-0.17	-5.10
	2016 – 2015 AzMERIT Match	-0.57	-0.43	-0.14	-4.20
	2016 – Math MC Match	-0.41	-0.24	-0.17	-5.10
G5M	2015	-0.09	-0.09	-0.01	-0.30
	2016 – 2014 AIMS Match	-0.06	-0.02	-0.04	-1.20
	2016 – 2015 AzMERIT Match	-0.16	-0.12	-0.03	-0.90
	2016 – Math MC Match	-0.07	-0.06	0.00	0.00
G6M	2015	0.07	0.01	0.07	2.10
	2016 – 2014 AIMS Match	-0.01	0.04	-0.05	-1.50
	2016 – 2015 AzMERIT Match	-0.09	-0.06	-0.03	-0.90
G7M	2015	0.15	0.07	0.08	2.40
	2016 – 2014 AIMS Match	0.18	0.07	0.11	3.30
	2016 – 2015 AzMERIT Match	0.11	-0.03	0.14	4.20
G8M	2015	0.43	0.32	0.11	3.30
	2016 – 2014 AIMS Match	0.56	0.55	0.00	0.00
	2016 – 2015 AzMERIT Match	0.47	0.47	0.01	0.30
Alg I	2015	0.29	0.23	0.05	1.50
	2016 – 2014 AIMS Match	0.64	0.51	0.13	3.90
	2016 – 2015 AzMERIT Match	0.72	0.57	0.15	4.50
Geo	2015	1.12	0.99	0.13	3.90
	2016 – 2014 AIMS Match	1.34	1.15	0.20	6.00
	2016 – 2015 AzMERIT Match	1.19	1.03	0.16	4.80
Alg II	2015	1.45	1.36	0.09	2.70
	2016 – 2014 AIMS Match	1.45	1.17	0.28	8.40
	2016 – 2015 AzMERIT Match	1.06	0.91	0.15	4.50

For grade 3 math assessment, as with grade 3 ELA, samples were matched based on student performance on the math multiple-choice items. Again this approach was applied in grades 4 and 5 to evaluate it against the other two methods, where the results indicated general convergence, indicating that items administered online were somewhat easier at grade 4 and no mode effect at grade 5. When applied at grade 3, a relatively large effect for mode was identified, indicating that items administered online were easier than on paper.

As with ELA, the identified mode effects varied across test administrations. The advantage of online over paper identified in 2016 was not observed in 2015. Likewise, observed effects of mode at grade 7 and for Algebra I and Algebra II in 2016 were not as pronounced in 2015, while effects of mode observed at grade 8 in 2015 were not observed in 2016. Thus, as with ELA, the effect of mode appears to be form specific and can be expected to vary across test administrations.

9.4.2 SCHOOL PERFORMANCE

In a separate approach to evaluating mode comparability, ADE implemented an investigation based on the spring 2015 operational test administration statewide (Scott, 2015). In her study, Scott (2015) first identified which Arizona schools elected to administer AzMERIT online and which on paper, and then examined the two samples of schools for any differences in performance on the spring 2014 paper administration of AIMS. The rationale in selecting school level analysis was based on schools having to choose only one of the two modes in which to assess all of their students. This increased level of matching was appropriate since the mode used by the student was, and continues to be, a school-based decision, rather than student based. Having found no difference in mean 2014 performance between the two groups, there would be no expectation for performance differences on AzMERIT except as a function of test administration mode. Following the spring 2015 administration of AzMERIT, ADE examined the performance of schools participating online and on paper, and again found performance on the AzMERIT to be comparable between the two sets of schools.

9.5 LINKING THE AZMERIT TO OTHER SCALES FOR PERFORMANCE COMPARISON

9.5.1 ESTABLISHING LINKAGES TO AIMS, SAGE, SMARTER BALANCED, PISA

To facilitate comparisons of Arizona achievement to other national and international benchmarks, a number of external linking sets were embedded in the 2015 AzMERIT field test slots. Arizona identified the locations of performance standards of other assessments systems on the AzMERIT scale; this information was used to inform panelists recommending performance standards for the AzMERIT.⁵⁶ The location of performance standards from the following assessments were identified on the AzMERIT scale:

- Smarter Balanced, by linking to AIR Core items on the Smarter Balanced scale,
- PISA, by embedding PISA items in the grade 10 ELA, Algebra I and Geometry EOC assessments
- historical Arizona performance by embedding AIMS items to link to the AIMS scale, and
- Utah's SAGE via common items in the operational test form.

Subsequent to calibration of the AzMERIT operational items and establishment of the reference scale, parameter estimates for those items were anchored to their reference values and all items administered in the embedded field test (EFT) blocks were calibrated under that constraint, placing parameter estimates for all field test and external linking item sets on the same AzMERIT scale defined by the operational item parameters. All external linking items had two sets of item parameters: a) external scale, and b) AzMERIT scale. To identify the location of external scale performance standards on the AzMERIT scale, AIR identified the linking constants necessary to

⁵⁶ Standard 5.23 – When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

transform item parameters from the external reference scale to the AzMERIT scale. Where the external scale was calibrated using the Rasch model, such as with AIMS, mean-sigma equating was used to identify the location of external performance standards on the AzMERIT scale. For external scales calibrated using more general IRT models, Stocking-Lord equating was used to identify the location of external scale performance standards on the AzMERIT scale.

In the context of standard setting, this procedure enabled ADE to identify a location in the AzMERIT Ordered Item Book (OIB) that represented a level of difficulty similar to a particular level in the external scale. For example, after finding the linking constant necessary to put the Smarter Balanced item parameters on the AzMERIT scale, it was possible to provide standard setting panelists with the location in the OIB that represents the level of difficulty comparable to each performance standard on the Smarter Balanced assessment.

9.5.2 IDENTIFYING THE LOCATION OF THE ACT COLLEGE-READY CUT ON AZMERIT

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the ACT college ready cuts were identified on the AzMERIT scale and provided to panelists during performance standards workshops in 2015. In order to identify the location of the ACT college ready cuts for the Grade 11 ELA and Algebra II AzMERIT End-of-Course assessments, a two-step approach was used to first identify the location of the ACT college ready benchmark on the AIMS scale, and then use the linkages between AIMS and AzMERIT to map the ACT college ready benchmark onto the AzMERIT scale(s). For this purpose, ADE provided ACT and AIMS scores for a recent cohort of students.

The sample used to investigate relationships between AzMERIT and ACT was based on records for students who took grade 11 ELA and Algebra II tests in spring 2015 and the ACT at an appropriate time for graduating in 2016. From among the full set of spring 2015 grade 11 ELA and Algebra II test takers, there are 58,888 (93%) and 32,945 (56%) grade 11 students, respectively. These records represent the target sample for the analyses reported in this study.

Because a large number of students did not take the ACT and the two subgroups differed systematically across demographic and achievement variables, the imputing approach is often employed to handle missing data in the analysis of the relationship between the AZMERIT scores and subsequent performance on the ACT. However, previous studies for Minnesota and Ohio showed that imputing or deleting the missing records did not impact the linkage identified between their graduation tests and the ACT test. In this study, the regression model that links the AzMERIT scale score to the ACT scale score was built based on the merged ACT and AzMERIT records. Then the validation set approach (training and testing set split) was used to estimate the variability of the model fit and to estimate the error rate of the model when applied in new previously unseen data.

ELA Examinees with missing ACT or AzMERIT scale scores were removed from the merged dataset. The ACT reading scale score for the remaining 25,977 students were regressed onto the applicable grade 11 ELA scale score and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2 , was identified as the best model to predict ACT reading from prior performance on the AzMERIT ELA test:

$$\hat{Y} = -290.65 + 0.12 * X_1 + 0.26 * X_2 - 2.35 * X_3 - 0.79 * X_4 + 0.57 * X_5 - 2.32 * X_6 - 1.79 * X_7 - 2.40 * X_8 - 1.82 * X_9 - 2.07 * X_{10}$$

where

\hat{Y} = ACT Reading Scale Score
 X1 = AZMERIT ELA Scale Score
 X2 = Female-Male Contrast
 X3 = American Indian-White Contrast
 X4 = Multi-ethnic Contrast
 X5 = Asian Contrast
 X6 = Hispanic-White Contrast
 X7 = African American-White Contrast
 X8 = Native Hawaiian-White Contrast
 X9 = Free and Reduced Lunch Contrast
 X10 = ELL Contrast

The overall model was statistically significant ($F(10, 20388) = 1704.70$, $p < .0001$; adjusted $R^2 = 0.46$). Table 1 shows the estimated model parameters. Application of this regression model indicates that an AzMERIT ELA scale score 2585 is associated with the ACT reading college ready cut score of 22.

Mathematics The records with missing ACT or AzMERIT scale scores were excluded from the analysis. Then the ACT mathematics scale scores for the remaining 13,777 students were regressed onto the applicable AzMERIT Algebra II test and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2 , was identified as the best model to predict ACT mathematics scores from prior performance on the AzMERIT Algebra II test:

$$\hat{Y} = -305.7 + 0.08 * X1 - 0.55 * X2 - 1.55 * X3 - 0.48 * X4 - 0.44 * X5 - 1.44 * X6 - 1.41 * X7 - 0.83 * X8 - 1.22 * X9 - 1.57 * X10$$

where

\hat{Y} = ACT Mathematics Scale Score
 X1 = AZMERIT Reading Scale Score
 X2 = Female-Male Contrast
 X3 = American Indian-White Contrast
 X4 = Multi-ethnic Contrast
 X5 = Asian Contrast
 X6 = Hispanic-White Contrast
 X7 = African American-White Contrast
 X8 = Native Hawaiian-White Contrast
 X9 = Free and Reduced Lunch Contrast
 X10 = ELL Contrast

The overall model was statistically significant ($F(10, 13768) = 1764.13$, $p < .0001$; adjusted $R^2 = 0.5$). Table 2 shows the estimated model parameters. Application of this regression model indicates that an AzMERIT mathematics score of 3727 is associated with the ACT mathematics college ready cut score of 22.

The validation set approach is a type of resampling method that estimates a model error rate by holding out a subset of the data from the fitting process (the testing dataset). The model is then built using the other set of observations (the training dataset). Then the model result is applied on the testing dataset in which we can then calculate the error. In summary, this general idea allows for the model to not overfit. In this study, the training dataset contained 50% randomly selected merged records and the testing dataset had the other 50% of students.

The multiple regression built by the training set yielded the same AzMERIT cut scores (ELA 2585, Math 3727) as the ones from the full data model. Then the predictive model was applied onto the testing set. The Root Mean Squared Error (RMSE) was calculated as the square root of the average squared errors found between the actual ACT score point and the model fitted values. Furthermore, we repeated this sampling and model fitting process 100 times to see how the RMSE varied across random samples. For ELA, the average RMSE was 5.03 and the standard deviation of the RMSE was 0.02 across the 100 replications. For mathematics, the average RMSE was 2.79 and the standard deviation was 0.02. The standard deviation of the RMSE was very small indicating that the sample selected for the modeling has no significant impact on the model fitting.

In addition, the equipercentile equating method was used to verify the linking between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT cut score 22 is 2585.72 for ELA and 3727.46 for mathematics. These cut scores are consistent with those identified using regression models

10. CONSTRUCTED-RESPONSE SCORING

The AzMERIT assessments in ELA and math utilize a variety of item types to assess students' mastery of the Arizona College and Career Ready Standards (ACCRS). ADE leverages AIR's item scoring technology to machine-score student responses to most items, including traditional selected-response (multiple choice) item types, and machine-scored constructed response (MSCR) items types. The MSCR item types are designed to capture and score a variety of response types, such as graphing, drawing or arranging graphic regions, selecting or rearranging sentences or phrases within passages, or entering equations or words, allowing AzMERIT items to assess a wide range of student knowledge and skills. In most cases, constructed response machine-scored items that are developed for online administration are adapted for paper and responses are captured in a format that allows machine-scoring.

In addition, some constructed response items are scored by human raters, also referred to as hand-scored. To support machine scoring of essay responses, a sample of essay responses was hand-scored through verification, and those responses and scores were used to develop the statistical scoring models used to score the remaining responses, and which will be used to score all essay responses in future test administrations. In addition, math assessments that were administered on paper included a small number of items that were scored by human raters. Generally, these were items that required students to produce an equation. The reading components of the ELA assessments, both online and paper, and the math assessments administered online are machine-scored in their entirety.

AIR partners with Measurement, Incorporated (MI) to fulfill all hand-scoring requirements. AIR provides the automated electronic scoring and MI provides all hand-scoring for the AzMERIT tests. This section describes the process for configuring and validating machine rubrics and the process for handscoring, including rules, descriptions of scorer training and systems used, and mechanisms for ensuring reliability and validity of item scores.

10.1 MACHINE SCORING

10.1.1 EXPLICIT RUBRICS

As part of the item development process for machine-scored item types which are scored with explicit rubrics, a rubric validation process was enacted to verify that rubrics are implemented as intended, and responses are scored correctly. This procedure is typically conducted following the initial administration of items, usually when the item is field-tested, and allows test developers to review the intent of the rubric versus the actual behavior. Actual student responses were reviewed by test development experts, along with resulting item scores, to ensure that the rubrics functioned as intended and awarded credit appropriately. Where necessary, test developers modified machine rubrics to address insufficiencies, automatically rescored student responses for the item, and repeating the process to finalize and approve the machine-scored rubrics. Test developers reviewed a strategic sample of responses, including responses where high achieving students scored poorly on the item, lower achieving students scored well on the item, and randomly selected responses from the population.

10.1.2 ESSAY AUTOSCORING

As part of the spring 2016 administration of AzMERIT, students in each grade were administered one of six writing tasks (three informational/explanatory and three persuasive/argumentative) in the writing component of each of

the ELA online assessments. Writing tasks were randomly administered to students. Random assignment of writing tasks helped ensure that writing tasks were calibrated on a representative sample of Arizona students, and served to reduce the effects of school clustering in the obtained samples, increasing the efficiency of the samples, and thereby yielding more precise item parameter estimates.

Two approaches were used to develop the statistical models that were used to score the essay responses. For AIRCore writing tasks that were administered online in the Florida field-test (grades 8-10), ADE adopted the scoring models generated from student responses in the Florida field test administration. Because the scoring models are based on semantic and syntactic features of the text that discriminate high versus low scoring essays as determined by human raters, the models are highly generalizable.

For the grades where scoring models did not already exist (3-7 and 11), an alternative approach was employed that allowed for autoscoring to be implemented as part of the spring 2016 essay scoring. Because the ELA window is split into separate writing and reading assessment windows, with the online writing window closing several weeks prior to close of the reading test administration, the dual window afforded an opportunity to build and implement the statistical scoring models in time to meet spring reporting timelines.

To facilitate development of the scoring models, MI conducted range finding, where possible, based on student responses from the Florida assessment. The range finding process is designed to calibrate a sample of responses for scorer training, qualification, and monitoring. Responses exemplifying each score point are identified and annotated for scorer training. Additional responses are identified for use in qualifying readers for scoring and for establishing validity sets that are used to monitor reader performance. Thus, for grades 4-7 which were included in the Florida field test, range finding activities to support AzMERIT rubric scoring were completed prior to the opening of the AzMERIT assessment window.

For the grade 3 and 11 assessments, which had not been previously administered, MI pulled a sample of essay responses following the first week of the testing window with which to conduct range finding activities. The development of training materials and training of raters followed immediately so that hand scoring could begin by the end of the fourth week of the testing window.

At the end of the second week of testing, AIR drew a random sample of 2,000 responses to each of the writing tasks administered at grades 3-7 and 11 for use in building the statistical scoring models. Those responses were routed to MI for hand-scoring. Each response was double scored, with any discrepancies routed for resolution scoring.

As hand-scoring activities were completed for each writing task, and scores were uploaded to AIR, work began to develop statistical scoring models for each rubric element, and to deploy those models to the test delivery system to score all remaining essay responses.⁵⁷

To develop the scoring models, the random sample of 2,000 responses was divided into a model building sample of 1,500 responses and a cross-validation sample of 500 responses. Model performance was evaluated on the cross-

⁵⁷ Standard 4.19 – When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

validation sample to ensure that model fit indices were not based on the model building sample, which may inflate fit indicators.

The statistical scoring models also yield an indicator of score confidence based on 1) responses with unusual features, and 2) responses scoring near rubric thresholds. For each model, a confidence threshold defined as two standard deviations below the mean confidence value for the responses in the cross validation sample was identified. Any scored response with a confidence value below the threshold was automatically routed to MI for verification scoring.

The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and “learned” from a training set. During training, these features are related to human scores through a statistical model. The resulting estimates complete a prediction equation that predicts how a human would score a response with the measured features. Statistical rubrics are, effectively, proxy measures. Although they can directly measure some aspects of writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not make direct measures of argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Further, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may *predict* whether a response to a factual item is correct—higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the “reason” that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

As noted, the engine employs a “training set,” a set of essay responses scored with maximally valid scores, which we obtain by having all responses double-scored by expert scorers and a thorough adjudication process for adjacent or discrepant scores. The quality of the human assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Approximately 1500 essay responses were selected at random from the set of scored essay responses to serve as the training set.

For each dimension in the rubric, the system estimates an appropriate statistical model relating the measures to the score assigned by humans. This model, along with its final parameter estimates, is used to generate a predicted or “proxy” score.

In addition to the training set, we draw an independent random sample of responses for cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are hand-scored, and agreement between human- and machine-assigned scores is examined. The cross-validation process ensures that the rubric generalizes across all responses and that the statistical model identified during training does not capitalize on peculiarities in the training set.

Exhibit 10.1.2.1 presents agreement indicators for the two initial human raters, and between the resolved human and statistical rubric score.⁵⁸ Indicators include percent exact agreement, Pearson’s correlation, a quadratic weighted kappa statistic, and the standardized mean difference between the scores. Although absolute values for evaluating statistics have been advanced (Condon, 2013; Higgins, 2013), the focus of these comparisons is

⁵⁸ Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

degradation of agreement when moving from human-human agreement to machine-human agreement. Agreement between human raters is an indicator of how reliably the responses can be scored by human raters. Since the statistical rubrics attempt to reproduce human assigned scores, evaluation of machine-human agreement is with respect to observed human-human agreement. Responses with poor human-human agreement will not be reliably scored by either humans or machines. Exhibit 10.1.2.2 presents the correlations among dimension scores for the summative and interim tests.

Exhibit 10.1.2.1 Summary of Human and Machine Scores for Spring 2016 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
3	13021	Conventions	2	2092	1.43	1.55	0.75	0.71	0.69	0.65	0.65	0.03	0.72	0.71	0.70	0.16
		Evidence	4		1.93	1.90	0.78	0.61	0.65	0.64	0.64	0.02	0.65	0.65	0.63	0.05
		Organization	4		1.93	2.00	0.76	0.66	0.66	0.67	0.67	0.00	0.67	0.66	0.65	0.10
3	13022	Conventions	2	2093	1.47	1.61	0.69	0.64	0.71	0.67	0.67	0.02	0.72	0.67	0.66	0.20
		Evidence	4		2.02	2.02	0.75	0.63	0.60	0.64	0.64	0.00	0.67	0.66	0.65	0.00
		Organization	4		2.12	2.10	0.74	0.66	0.64	0.68	0.68	0.00	0.68	0.65	0.64	0.03
3	13023	Conventions	2	2090	1.51	1.57	0.72	0.65	0.70	0.65	0.65	0.01	0.73	0.67	0.66	0.09
		Evidence	4		1.89	1.93	0.74	0.65	0.62	0.60	0.60	0.03	0.68	0.68	0.67	0.05
		Organization	4		1.95	1.92	0.77	0.61	0.64	0.66	0.66	0.02	0.68	0.66	0.64	0.04
3	13024	Conventions	2	2096	1.44	1.53	0.70	0.67	0.71	0.66	0.66	0.01	0.76	0.68	0.67	0.13
		Evidence	4		1.93	1.90	0.76	0.64	0.63	0.64	0.64	0.04	0.64	0.63	0.62	0.04
		Organization	4		1.96	1.96	0.80	0.65	0.63	0.66	0.66	0.05	0.64	0.64	0.63	0.00
3	13025	Conventions	2	2093	1.37	1.48	0.76	0.70	0.66	0.58	0.58	0.01	0.71	0.67	0.66	0.15
		Evidence	4		1.94	1.97	0.72	0.61	0.63	0.65	0.65	0.00	0.69	0.65	0.64	0.04
		Organization	4		1.92	1.86	0.82	0.71	0.61	0.64	0.64	0.01	0.64	0.68	0.68	0.08
3	13026	Conventions	2	2090	1.45	1.55	0.73	0.67	0.71	0.66	0.66	0.00	0.75	0.68	0.67	0.15
		Evidence	4		1.94	1.94	0.74	0.68	0.66	0.68	0.68	0.04	0.72	0.71	0.71	0.01
		Organization	4		2.04	2.02	0.80	0.71	0.64	0.68	0.68	0.05	0.63	0.65	0.65	0.03
4	13094	Conventions	2	2095	0.95	0.95	0.75	0.68	0.66	0.67	0.67	0.00	0.65	0.65	0.65	0.01
		Evidence	4		1.30	1.27	0.47	0.47	0.77	0.52	0.52	0.00	0.82	0.58	0.58	0.08
		Organization	4		1.40	1.34	0.51	0.49	0.74	0.56	0.56	0.01	0.83	0.66	0.66	0.11
4	13095	Conventions	2	2096	1.17	1.17	0.67	0.63	0.64	0.62	0.62	0.01	0.67	0.59	0.59	0.01
		Evidence	4		1.35	1.24	0.53	0.45	0.75	0.57	0.57	0.00	0.81	0.63	0.60	0.22
		Organization	4		1.54	1.51	0.59	0.54	0.71	0.59	0.59	0.03	0.73	0.56	0.56	0.06
4	13118	Conventions	2	2096	1.15	1.16	0.71	0.65	0.64	0.60	0.60	0.01	0.67	0.63	0.63	0.01
		Evidence	4		1.33	1.29	0.49	0.48	0.76	0.55	0.55	0.01	0.84	0.64	0.64	0.07
		Organization	4		1.56	1.53	0.61	0.56	0.71	0.59	0.59	0.03	0.77	0.67	0.67	0.04
4	13119	Conventions	2	2094	1.15	1.19	0.72	0.63	0.66	0.64	0.64	0.02	0.66	0.63	0.63	0.06
		Evidence	4*		1.38	1.30	0.54	0.49	0.73	0.53	0.53	0.05	0.77	0.57	0.56	0.14

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
		Organization	4		1.51	1.46	0.60	0.52	0.72	0.60	0.60	0.01	0.75	0.60	0.60	0.10
4	13120	Conventions	2	2091	1.05	1.09	0.70	0.67	0.67	0.66	0.66	0.02	0.68	0.64	0.64	0.06
		Evidence	4		1.28	1.20	0.49	0.42	0.77	0.54	0.54	0.04	0.85	0.65	0.63	0.17
		Organization	4		1.49	1.43	0.58	0.53	0.74	0.63	0.63	0.03	0.79	0.65	0.64	0.11
4	13121	Conventions	2	2096	1.10	1.08	0.69	0.59	0.67	0.65	0.65	0.03	0.68	0.61	0.60	0.02
		Evidence	4*		1.34	1.27	0.54	0.49	0.77	0.60	0.60	0.03	0.81	0.65	0.64	0.14
		Organization	4*		1.53	1.45	0.58	0.54	0.72	0.61	0.61	0.03	0.74	0.59	0.59	0.13
5	13236	Conventions	2	2099	1.41	1.57	0.67	0.62	0.74	0.69	0.69	0.02	0.76	0.69	0.67	0.25
		Evidence	4		1.81	1.79	0.58	0.52	0.71	0.59	0.59	0.01	0.79	0.64	0.64	0.03
		Organization	4		1.92	1.88	0.68	0.58	0.70	0.65	0.65	0.03	0.73	0.67	0.66	0.05
5	13237	Conventions	2	2095	1.30	1.40	0.74	0.67	0.73	0.72	0.71	0.04	0.73	0.69	0.68	0.13
		Evidence	4		1.59	1.53	0.60	0.53	0.73	0.61	0.61	0.04	0.76	0.62	0.62	0.09
		Organization	4		1.75	1.75	0.66	0.57	0.72	0.66	0.66	0.01	0.72	0.64	0.64	0.01
5	13238	Conventions	2	2099	1.47	1.51	0.62	0.61	0.72	0.65	0.65	0.00	0.75	0.65	0.64	0.06
		Evidence	4		1.87	1.88	0.64	0.53	0.69	0.63	0.63	0.01	0.75	0.63	0.62	0.02
		Organization	4		1.95	1.99	0.68	0.56	0.70	0.65	0.65	0.01	0.74	0.62	0.61	0.06
5	13239	Conventions	2	2095	1.41	1.51	0.69	0.60	0.73	0.66	0.66	0.02	0.75	0.68	0.67	0.15
		Evidence	4		1.67	1.67	0.62	0.56	0.65	0.56	0.56	0.02	0.74	0.63	0.63	0.01
		Organization	4		1.92	1.93	0.64	0.52	0.71	0.65	0.65	0.02	0.76	0.63	0.61	0.03
5	13246	Conventions	2	2093	1.36	1.45	0.68	0.65	0.72	0.68	0.68	0.01	0.73	0.69	0.69	0.13
		Evidence	4		1.54	1.58	0.57	0.54	0.72	0.59	0.59	0.03	0.77	0.61	0.60	0.07
		Organization	4		1.81	1.82	0.66	0.57	0.71	0.65	0.65	0.01	0.73	0.64	0.64	0.02
5	13247	Conventions	2	2097	1.38	1.43	0.68	0.63	0.72	0.67	0.67	0.01	0.75	0.71	0.71	0.07
		Evidence	4		1.77	1.80	0.67	0.59	0.65	0.60	0.60	0.02	0.73	0.65	0.65	0.05
		Organization	4		2.00	1.97	0.69	0.57	0.69	0.66	0.66	0.02	0.72	0.63	0.62	0.04
6	13304	Conventions	2	2097	1.43	1.52	0.67	0.62	0.67	0.57	0.57	0.02	0.76	0.72	0.71	0.14
		Evidence	4		1.74	1.75	0.65	0.61	0.63	0.57	0.56	0.04	0.74	0.66	0.66	0.02
		Organization	4		1.89	1.86	0.74	0.65	0.62	0.61	0.61	0.01	0.69	0.68	0.67	0.04
6	13305	Conventions	2	2095	1.45	1.59	0.68	0.61	0.66	0.57	0.57	0.00	0.76	0.69	0.67	0.22
		Evidence	4		1.53	1.43	0.60	0.55	0.70	0.58	0.58	0.01	0.74	0.61	0.60	0.17
		Organization	4		1.62	1.60	0.68	0.62	0.65	0.59	0.59	0.01	0.70	0.62	0.62	0.02

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
6	13306	Conventions	2	2097	1.47	1.54	0.69	0.64	0.71	0.64	0.64	0.02	0.75	0.68	0.67	0.11
		Evidence	4		1.67	1.63	0.64	0.57	0.65	0.55	0.55	0.00	0.71	0.61	0.60	0.08
		Organization	4		1.85	1.80	0.69	0.61	0.64	0.62	0.61	0.04	0.71	0.67	0.66	0.07
6	13307	Conventions	2	2095	1.36	1.42	0.69	0.65	0.66	0.64	0.64	0.04	0.72	0.68	0.68	0.09
		Evidence	4		1.54	1.52	0.68	0.65	0.67	0.62	0.62	0.04	0.72	0.64	0.64	0.03
		Organization	4		1.78	1.80	0.74	0.63	0.62	0.61	0.61	0.02	0.68	0.65	0.64	0.02
6	13308	Conventions	2	2097	1.41	1.50	0.67	0.62	0.62	0.54	0.54	0.06	0.74	0.68	0.67	0.13
		Evidence	4		1.46	1.37	0.62	0.57	0.69	0.57	0.57	0.03	0.71	0.59	0.58	0.15
		Organization	4		1.64	1.57	0.69	0.62	0.63	0.60	0.60	0.03	0.71	0.66	0.65	0.10
6	13309	Conventions	2	2093	1.39	1.48	0.65	0.56	0.68	0.58	0.58	0.06	0.76	0.68	0.67	0.15
		Evidence	4		1.69	1.60	0.73	0.67	0.65	0.59	0.59	0.02	0.72	0.71	0.70	0.13
		Organization	4		1.84	1.83	0.78	0.69	0.61	0.62	0.62	0.01	0.70	0.71	0.71	0.01
7	13400	Conventions	2	2082	1.35	1.45	0.66	0.63	0.70	0.67	0.67	0.02	0.74	0.70	0.69	0.14
		Evidence	4		1.84	1.83	0.61	0.53	0.66	0.60	0.60	0.07	0.77	0.65	0.65	0.03
		Organization	4		1.92	1.90	0.64	0.54	0.65	0.62	0.62	0.02	0.74	0.61	0.60	0.03
7	13401	Conventions	2	2084	1.65	1.72	0.56	0.49	0.79	0.62	0.62	0.04	0.80	0.64	0.63	0.14
		Evidence	4		1.86	1.87	0.58	0.50	0.72	0.63	0.63	0.01	0.79	0.64	0.63	0.03
		Organization	4		2.00	2.02	0.54	0.48	0.73	0.59	0.59	0.02	0.83	0.66	0.65	0.05
7	13402	Conventions	2	2088	1.49	1.55	0.63	0.62	0.69	0.60	0.60	0.03	0.75	0.67	0.67	0.10
		Evidence	4		1.83	1.87	0.51	0.43	0.73	0.59	0.59	0.04	0.88	0.74	0.72	0.08
		Organization	4		1.91	1.93	0.59	0.50	0.70	0.61	0.61	0.01	0.80	0.66	0.65	0.02
7	13403	Conventions	2	2085	1.56	1.62	0.57	0.55	0.77	0.61	0.61	0.03	0.81	0.70	0.70	0.11
		Evidence	4		1.65	1.58	0.60	0.57	0.73	0.66	0.66	0.00	0.80	0.72	0.72	0.12
		Organization	4		1.75	1.75	0.64	0.56	0.68	0.61	0.61	0.02	0.78	0.69	0.69	0.00
7	13405	Conventions	2	2093	1.46	1.48	0.61	0.62	0.75	0.63	0.63	0.03	0.77	0.68	0.68	0.02
		Evidence	4		1.63	1.66	0.59	0.62	0.74	0.70	0.70	0.03	0.79	0.72	0.72	0.04
		Organization	4		1.83	1.80	0.62	0.56	0.73	0.68	0.68	0.03	0.79	0.70	0.69	0.05
7	13406	Conventions	2	2090	1.44	1.47	0.62	0.58	0.72	0.63	0.63	0.01	0.76	0.67	0.67	0.05
		Evidence	4		1.80	1.81	0.54	0.46	0.73	0.58	0.58	0.03	0.79	0.58	0.57	0.01
		Organization	4		1.92	1.89	0.56	0.49	0.71	0.59	0.59	0.01	0.79	0.62	0.62	0.06
8	13437	Conventions	2	2391	1.47	1.53	0.68	0.63	0.74	0.69	0.69	0.03	0.77	0.72	0.72	0.09

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
		Evidence	4		1.91	1.89	0.68	0.61	0.76	0.75	0.75	0.01	0.75	0.70	0.70	0.02
		Organization	4		2.05	2.01	0.77	0.69	0.73	0.75	0.75	0.01	0.75	0.75	0.74	0.06
8	13438	Conventions	2	2631	2.01	1.95	0.77	0.71	0.79	0.70	0.70	0.01	0.73	0.75	0.74	0.08
		Evidence	4		2.11	2.08	0.80	0.76	0.77	0.78	0.78	0.00	0.71	0.75	0.74	0.04
		Organization	4		1.55	1.57	0.63	0.59	0.73	0.76	0.76	0.01	0.79	0.72	0.72	0.03
8	13439	Conventions	2	2548	1.57	1.67	0.63	0.55	0.78	0.67	0.67	0.02	0.83	0.76	0.74	0.17
		Evidence	4		2.05	2.07	0.72	0.60	0.73	0.74	0.74	0.01	0.73	0.70	0.69	0.03
		Organization	4		2.15	2.16	0.81	0.73	0.71	0.74	0.74	0.01	0.69	0.72	0.71	0.01
8	13452	Conventions	2	2491	1.61	1.65	0.58	0.54	0.79	0.67	0.67	0.02	0.80	0.68	0.68	0.06
		Evidence	4		2.07	2.06	0.75	0.64	0.77	0.77	0.77	0.01	0.74	0.73	0.72	0.02
		Organization	4		2.20	2.18	0.76	0.67	0.74	0.75	0.75	0.01	0.75	0.76	0.76	0.03
8	13453	Conventions	2	2538	1.53	1.57	0.64	0.60	0.76	0.68	0.68	0.01	0.78	0.71	0.71	0.06
		Evidence	4		1.99	1.99	0.78	0.74	0.76	0.78	0.78	0.01	0.73	0.76	0.76	0.00
		Organization	4		2.14	2.12	0.79	0.73	0.75	0.79	0.79	0.02	0.74	0.77	0.77	0.03
8	13454	Conventions	2	2544	1.56	1.58	0.61	0.56	0.77	0.68	0.68	0.01	0.78	0.68	0.68	0.03
		Evidence	4		1.99	1.91	0.74	0.67	0.77	0.77	0.77	0.01	0.74	0.73	0.73	0.10
		Organization	4		2.04	2.06	0.76	0.74	0.75	0.77	0.77	0.01	0.74	0.76	0.76	0.02
9	13554	Conventions	2	2751	1.61	1.68	0.59	0.55	0.81	0.71	0.71	0.02	0.80	0.69	0.68	0.13
		Evidence	4		1.89	1.92	0.62	0.53	0.82	0.76	0.76	0.01	0.79	0.68	0.67	0.04
		Organization	4		2.02	2.03	0.65	0.60	0.79	0.76	0.76	0.02	0.80	0.74	0.73	0.01
9	13555	Conventions	2	2853	1.58	1.68	0.63	0.57	0.81	0.74	0.74	0.02	0.81	0.76	0.74	0.17
		Evidence	4		1.88	1.90	0.66	0.59	0.81	0.77	0.77	0.02	0.79	0.72	0.72	0.03
		Organization	4		2.02	1.99	0.70	0.66	0.80	0.79	0.79	0.01	0.80	0.78	0.78	0.05
9	13556	Conventions	2	1469	1.66	1.72	0.57	0.56	0.80	0.69	0.69	0.02	0.81	0.71	0.71	0.11
		Evidence	4		1.86	1.90	0.64	0.60	0.79	0.76	0.76	0.01	0.78	0.71	0.71	0.08
		Organization	4		2.00	2.01	0.70	0.62	0.77	0.77	0.77	0.02	0.77	0.74	0.73	0.01
9	13557	Conventions	2	2815	1.54	1.58	0.65	0.60	0.79	0.73	0.73	0.00	0.78	0.71	0.70	0.06
		Evidence	4		1.82	1.85	0.57	0.53	0.83	0.78	0.78	0.01	0.83	0.72	0.72	0.05
		Organization	4		1.99	1.97	0.69	0.65	0.79	0.78	0.78	0.01	0.80	0.77	0.77	0.02
9	13565	Conventions	2	2869	1.52	1.56	0.62	0.61	0.80	0.75	0.75	0.02	0.78	0.71	0.71	0.06
		Evidence	4		1.92	1.92	0.67	0.60	0.81	0.80	0.80	0.03	0.79	0.74	0.73	0.01

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
		Organization	4		2.11	2.11	0.72	0.66	0.79	0.80	0.80	0.01	0.78	0.76	0.76	0.00
9	13566	Conventions	2	2852	1.54	1.59	0.63	0.60	0.81	0.76	0.76	0.03	0.81	0.76	0.76	0.09
		Evidence	4		1.93	1.93	0.62	0.54	0.84	0.80	0.80	0.00	0.82	0.74	0.73	0.01
		Organization	4		2.08	2.09	0.67	0.63	0.79	0.79	0.79	0.02	0.82	0.77	0.77	0.00
10	13635	Conventions	2	2436	1.61	1.65	0.55	0.53	0.71	0.60	0.60	0.02	0.77	0.61	0.61	0.07
		Evidence	4		2.04	2.08	0.77	0.71	0.69	0.73	0.73	0.01	0.75	0.76	0.76	0.05
		Organization	4		2.25	2.26	0.76	0.69	0.70	0.73	0.73	0.04	0.72	0.73	0.72	0.02
10	13636	Conventions	2	2344	1.69	1.78	0.49	0.45	0.72	0.58	0.57	0.01	0.83	0.63	0.62	0.19
		Evidence	4		1.99	1.96	0.74	0.66	0.74	0.73	0.73	0.00	0.76	0.76	0.76	0.04
		Organization	4		2.06	2.08	0.75	0.72	0.72	0.74	0.74	0.00	0.79	0.81	0.81	0.03
10	13637	Conventions	2	1314	1.58	1.65	0.60	0.53	0.70	0.59	0.59	0.02	0.76	0.62	0.61	0.11
		Evidence	4		1.89	1.88	0.70	0.66	0.76	0.72	0.72	0.01	0.77	0.75	0.75	0.02
		Organization	4		2.06	2.04	0.68	0.60	0.75	0.72	0.72	0.03	0.76	0.69	0.69	0.03
10	13638	Conventions	2	2475	1.62	1.69	0.54	0.48	0.70	0.57	0.56	0.03	0.79	0.60	0.59	0.14
		Evidence	4		1.99	2.00	0.74	0.64	0.71	0.71	0.71	0.02	0.74	0.74	0.73	0.02
		Organization	4		2.12	2.14	0.76	0.70	0.69	0.71	0.71	0.01	0.77	0.79	0.78	0.03
10	13639	Conventions	2	2306	1.66	1.73	0.53	0.47	0.70	0.56	0.56	0.03	0.80	0.60	0.59	0.14
		Evidence	4		1.97	1.96	0.72	0.62	0.72	0.69	0.69	0.04	0.74	0.72	0.71	0.01
		Organization	4		2.10	2.14	0.71	0.69	0.70	0.70	0.70	0.03	0.78	0.75	0.75	0.05
10	13640	Conventions	2	2399	1.68	1.71	0.52	0.51	0.74	0.61	0.61	0.03	0.81	0.65	0.65	0.05
		Evidence	4		2.14	2.11	0.81	0.71	0.69	0.71	0.71	0.02	0.71	0.75	0.74	0.04
		Organization	4		2.26	2.29	0.79	0.72	0.67	0.71	0.71	0.04	0.73	0.76	0.76	0.04
11	13720	Conventions	2	2091	1.56	1.56	0.61	0.58	0.74	0.65	0.65	0.03	0.72	0.58	0.58	0.00
		Evidence	4		1.91	1.88	0.77	0.72	0.62	0.67	0.67	0.00	0.70	0.72	0.72	0.05
		Organization	4		2.15	2.15	0.77	0.68	0.67	0.71	0.71	0.01	0.73	0.73	0.72	0.00
11	13721	Conventions	2	2090	1.56	1.62	0.61	0.57	0.74	0.62	0.62	0.01	0.74	0.62	0.61	0.10
		Evidence	4		2.18	2.16	0.77	0.66	0.60	0.66	0.66	0.03	0.68	0.69	0.68	0.03
		Organization	4		2.36	2.35	0.71	0.64	0.66	0.67	0.66	0.03	0.76	0.74	0.73	0.01
11	13722	Conventions	2	2090	1.58	1.65	0.60	0.57	0.78	0.68	0.68	0.01	0.81	0.74	0.73	0.12
		Evidence	4		2.18	2.20	0.80	0.70	0.62	0.67	0.67	0.02	0.70	0.74	0.73	0.03
		Organization	4		2.38	2.36	0.73	0.68	0.66	0.70	0.70	0.04	0.76	0.75	0.75	0.03

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ^*	SMD*	% Exact	Pearson r	Weighted κ^*	SMD*
11	13723	Conventions	2	2095	1.60	1.63	0.59	0.57	0.75	0.61	0.61	0.01	0.77	0.65	0.65	0.05
		Evidence	4		2.24	2.24	0.83	0.74	0.62	0.70	0.70	0.02	0.70	0.74	0.73	0.00
		Organization	4		2.47	2.47	0.74	0.68	0.64	0.69	0.69	0.00	0.74	0.74	0.73	0.01
11	13724	Conventions	2	2089	1.60	1.61	0.57	0.54	0.73	0.61	0.61	0.03	0.79	0.66	0.66	0.03
		Evidence	4		2.24	2.27	0.79	0.74	0.63	0.68	0.68	0.00	0.74	0.76	0.75	0.04
		Organization	4		2.28	2.32	0.75	0.68	0.64	0.66	0.66	0.01	0.75	0.76	0.75	0.06
11	13725	Conventions	2	2085	1.45	1.50	0.67	0.63	0.71	0.64	0.64	0.04	0.78	0.74	0.73	0.08
		Evidence	4		2.21	2.26	0.83	0.78	0.64	0.72	0.72	0.02	0.74	0.79	0.79	0.05
		Organization	4		2.36	2.34	0.81	0.71	0.66	0.73	0.73	0.00	0.74	0.78	0.77	0.02

Note. Weighted K = Quadratic weighted kappa; SMD = Standardized Mean Difference

*For asterisked items, no 4-point responses were identified in the training set, so at present statistical models for these items can only assign up to three points.

Exhibit 10.1.2.2. Summary of Dimension Intercorrelations for Spring 2016 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
3	13021	Conventions	2	2092		
		Evidence	4		0.61	
		Organization	4		0.60	0.79
3	13022	Conventions	2	2093		
		Evidence	4		0.65	
		Organization	4		0.63	0.87
3	13023	Conventions	2	2090		
		Evidence	4		0.67	
		Organization	4		0.72	0.86
3	13024	Conventions	2	2096		
		Evidence	4		0.72	
		Organization	4		0.61	0.89
3	13025	Conventions	2	2093		
		Evidence	4		0.65	
		Organization	4		0.64	0.83
3	13026	Conventions	2	2090		
		Evidence	4		0.66	
		Organization	4		0.61	0.88
4	13094	Conventions	2	2095		
		Evidence	4		0.63	
		Organization	4		0.62	0.72
4	13095	Conventions	2	2096		
		Evidence	4		0.55	
		Organization	4		0.67	0.64
4	13118	Conventions	2	2096		
		Evidence	4		0.51	
		Organization	4		0.67	0.61
4	13119	Conventions	2	2094		
		Evidence	4		0.52	
		Organization	4		0.64	0.63
4	13120	Conventions	2	2091		
		Evidence	4		0.55	
		Organization	4		0.72	0.70
4	13121	Conventions	2	2096		
		Evidence	4		0.49	
		Organization	4		0.57	0.66
5	13236	Conventions	2	2099		
		Evidence	4		0.52	
		Organization	4		0.57	0.81
5	13237	Conventions	2	2095		
		Evidence	4		0.59	
		Organization	4		0.63	0.71
5	13238	Conventions	2	2099		
		Evidence	4		0.57	
		Organization	4		0.53	0.78
5	13239	Conventions	2	2095		
		Evidence	4		0.63	
		Organization	4		0.66	0.77

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
5	13246	Conventions Evidence Organization	2 4 4	2093	0.64 0.59	0.68
5	13247	Conventions Evidence Organization	2 4 4	2097	0.58 0.46	0.86
6	13304	Conventions Evidence Organization	2 4 4	2097	0.72 0.68	0.90
6	13305	Conventions Evidence Organization	2 4 4	2095	0.53 0.62	0.77
6	13306	Conventions Evidence Organization	2 4 4	2097	0.66 0.69	0.76
6	13307	Conventions Evidence Organization	2 4 4	2095	0.69 0.68	0.72
6	13308	Conventions Evidence Organization	2 4 4	2097	0.42 0.62	0.76
6	13309	Conventions Evidence Organization	2 4 4	2093	0.74 0.71	0.77
7	13400	Conventions Evidence Organization	2 4 4	2082	0.63 0.65	0.73
7	13401	Conventions Evidence Organization	2 4 4	2084	0.66 0.49	0.80
7	13402	Conventions Evidence Organization	2 4 4	2088	0.64 0.64	0.87
7	13403	Conventions Evidence Organization	2 4 4	2085	0.57 0.71	0.63
7	13405	Conventions Evidence Organization	2 4 4	2093	0.58 0.62	0.76
7	13406	Conventions Evidence Organization	2 4 4	2090	0.61 0.62	0.69
8	13437	Conventions Evidence Organization	2 4 4	2391	0.55 0.42	0.85
8	13438	Conventions Evidence Organization	2 4 4	2631	0.89 0.61	0.49
8	13439	Conventions Evidence	2 4	2548	0.47	

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
		Organization	4		0.51	0.83
8	13452	Conventions	2	2491		
		Evidence	4		0.56	
		Organization	4		0.61	0.86
8	13453	Conventions	2	2538		
		Evidence	4		0.60	
		Organization	4		0.61	0.85
8	13454	Conventions	2	2544		
		Evidence	4		0.51	
		Organization	4		0.53	0.80
9	13554	Conventions	2	2751		
		Evidence	4		0.47	
		Organization	4		0.50	0.79
9	13555	Conventions	2	2853		
		Evidence	4		0.72	
		Organization	4		0.67	0.81
9	13556	Conventions	2	1469		
		Evidence	4		0.51	
		Organization	4		0.57	0.82
9	13557	Conventions	2	2815		
		Evidence	4		0.39	
		Organization	4		0.49	0.76
9	13565	Conventions	2	2869		
		Evidence	4		0.63	
		Organization	4		0.59	0.81
9	13566	Conventions	2	2852		
		Evidence	4		0.58	
		Organization	4		0.59	0.78
10	13635	Conventions	2	2436		
		Evidence	4		0.52	
		Organization	4		0.51	0.87
10	13636	Conventions	2	2344		
		Evidence	4		0.46	
		Organization	4		0.51	0.79
10	13637	Conventions	2	1314		
		Evidence	4		0.51	
		Organization	4		0.50	0.81
10	13638	Conventions	2	2475		
		Evidence	4		0.51	
		Organization	4		0.57	0.83
10	13639	Conventions	2	2306		
		Evidence	4		0.49	
		Organization	4		0.58	0.80
10	13640	Conventions	2	2399		
		Evidence	4		0.43	
		Organization	4		0.64	0.82
11	13720	Conventions	2	2091		
		Evidence	4		0.61	
		Organization	4		0.51	0.79
11	13721	Conventions	2	2090		

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
		Evidence	4		0.61	
		Organization	4		0.54	0.79
11	13722	Conventions	2	2090		
		Evidence	4		0.63	
		Organization	4		0.63	0.80
11	13723	Conventions	2	2095		
		Evidence	4		0.59	
		Organization	4		0.63	0.77
11	13724	Conventions	2	2089		
		Evidence	4		0.70	
		Organization	4		0.66	0.88
11	13725	Conventions	2	2085		
		Evidence	4		0.65	
		Organization	4		0.62	0.87

10.2 HAND-SCORING

Hand scoring of online essay responses for statistical model building, as well as hand scoring of all essay responses from paper test administrations were routed to MI for scoring. As noted in section 10.1, the sample of essay responses selected for statistical model building was independently scored by two readers. Any response assigned discrepant scores were routed for resolution scoring by a scoring trainer. In addition, all essay responses captured from paper test administrations were hand scored, with ten percent of all paper responses receiving a second reading (Reader 2) for the purpose of monitoring and maintaining sufficient inter-rater reliability, as discussed below. For ELA hand scoring, where scores from Reader 1 and Reader 2 were not in adjacent agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score was based on the resolution score, when present, or else on the initial read. For math hand scoring, where scores from Reader 1 and Reader 2 were not in exact agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score for math was based on the resolution score, when present, or else on the initial read.

10.2.1 HANDSCORING PROCESS

MI's hand-scoring efforts are managed via the Virtual Scoring Center VSC™ software, which is composed of two primary subsystems: VSC Capture™ and VSC Score™. Images of student responses to open ended items were sent to VSC Score™, which is a web-based environment for scoring constructed-response items by scorers working in an online environment. VSC Score is a secure, centrally administered environment used by site-based scorers. The interface enabled scorers to evaluate constructed response items and writing assessments from images. VSC Score has the following capabilities:

- Defining scorer roles and qualifications based on training, security requirements, or prior history
- Managing and randomly routing scorers' responses that require second readings in a double-blind manner
- Allowing project leaders to spot-check scorers, monitor reliability, and offer feedback
- Allowing scorers to flag responses for a variety of reasons (unusual approaches, nonscorable issues, etc.)
- Generating status reports at project milestones (such as percent of items scored)
- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
- Accommodating paper-based scores when images are of insufficient quality

- Outputting data easily into MI's score reporting applications

Paper-pencil tests were scanned into VSC. The images were displayed to trained and qualified scorers who scored the images online. Scorers had access only to those items for which they had been qualified to score. Online assessment responses were also converted into images and displayed in an identical manner to paper-pencil student responses using the same VSC scoring application.

When logging onto VSC Score, scorers were presented with a scoring set, which is the images-scoring equivalent of a physical packet of student responses. The scoring set was generated by randomly selecting student responses from the pool of non-scored student responses. The resultant set of responses was checked out to the scorer. The images they received had no demographic information on them. The scorer did not know the name, sex, school, or location of the student whose item was being scored. The scorer evaluated the first response, entered the score by clicking the appropriate values on the scoring toolbar, and clicked the Submit button. For multi-page responses, a scorer had to view each page of the response before a score was entered. Once the Submit button was clicked, the system recorded the score and the next response in the scoring set appeared for the scorer to score and submit. This process continued until all responses in the set had been scored.

When a scorer had a question about a response, he or she transferred the image (along with a virtual note including the question and/or comments) from the current scoring set to a review set assigned to a team leader or the scoring director. The team leader or scoring director submitted the appropriate score or returned the response to the scorer with comments. This procedure was used whenever a scorer had scoring concerns or found nonscorable responses (NSR) or responses requiring condition codes. Previously, condition codes were assigned to student responses by scoring leadership per Arizona specifications, such as a code noting that the response was left blank, the response was undecipherable or illegible, non-English, and so on. Condition codes other than blank were then recoded to the lowest score for each dimension for ability estimation. Because the statistical scoring engine cannot assign condition codes, all non-blank responses were assigned a rubric score directly, with responses that would otherwise have received a non-blank condition code being assigned the lowest score point for each dimension.

After scoring all of the responses in a set, the scorer reviewed any of the responses and modified the scores before committing them to the system. Once the scores had been committed, the set was checked in and responses are routed to other scorers as necessary. Regardless of the specific requirements, however, student responses were not marked as complete until the requisite number of independent scorers had scored the response.

VSC prioritized the available responses in the queue to make sure that the newer responses were placed toward the back of the queue.

10.2.2 HAND-SCORING QUALITY CONTROL

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students. The requirements for double scorings are defined to VSC at setup time. MI assumed a double-blind scoring rate of 10% for both the essays and math constructed responses.

10.2.3 HAND-SCORING RELIABILITY AND VALIDITY

MI uses a two-pronged approach to construct the scoring teams for AzMERIT. First, the scoring leadership recruits qualified, experienced scorers who have successfully scored large-scale assessments for MI, and therefore have

experience understanding the approach to scoring. To ensure reliable and valid hand-scores, MI puts scoring directors, team leaders, and scorers through a rigorous screening and training process.⁵⁹

Scoring directors, team leaders, and scorers are hired for AzMERIT based on experience and performance. Potential new scorers are given a comprehensive content screening for reading and math. This screening is used to identify potential scorers' aptitude for content area and grade level as well as their reading comprehension and deductive reasoning skills, which are directly related to what they may be scoring. In addition to writing an extemporaneous essay, new hires are required to read a passage and answer questions pertaining to that passage, proofread a sample essay for writing conventions, and solve a series of math problems. The results determine grade and content area placement if a scorer is to be offered a position on a project. New scorers are selected based on their scores on MI's content screening assessment given for language arts and math projects, the quality of their interview, their work history, and the references provided. The actual qualification for the scorers occurs at the end of training. In addition, the scorers are provided with ongoing validation that they are providing the state with consistency in their scoring through the use of validation sets that are incorporated into the ongoing live scoring.

All of the Arizona training materials provided for the initial operational ELA scoring were scoring guides composed of anchor responses as well as training, qualifying, and recalibration sets approved for use by the state as a result of approval of existing documentation from AIR's Item Tracking System (ITS), which is the repository for all item attributes, including scoring rubrics. In subsequent years, new items approved from the previous year's field test will be incorporated based on the materials used during the field test scoring. All materials and selected sets were submitted to Arizona for approval.⁶⁰

MI's scoring directors ensured ELA scoring guides had detailed annotations to explain how the scoring criteria are to be applied to each response's specific features and why it should be assigned a particular score. The approach was to focus on the precise scoring rationale, which helped scorers define the lines between score points. All scoring guides and other training materials were presented to Arizona for review and approval prior to the start of scoring.

Training sets and qualifying sets consisted of items that are most representative of the type that will be scored. MI scoring leadership selected these responses and provided them to Arizona for approval prior to their use. The training and qualifying sets contained examples of responses from all score points arranged in random score-point order. MI created an appropriate number of training sets and qualifying sets based on the complexity of the item. Essay questions were more complex than single-point math items. The sets were designed to help the scorers learn to apply the criteria illustrated in the scoring guide, ensure that the scorers become familiar with the process of scoring student responses, and assess the scorers' understanding of the scoring criteria before they are allowed to begin live scoring. MI worked with Arizona to finalize the number of training and qualifying sets for each item and determine the appropriate qualifying percentage. All scoring decisions and supplemental responses were submitted more than one month before the start of scoring for review and approval by the state.

MI's scoring directors trained both new and experienced scorers within the scoring rooms, giving detailed explanations of all training materials.

⁵⁹ Standard 4.20 – The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

⁶⁰ Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

MI's online training interface allowed observers from ADE to witness training in real-time. Through the use of TurboMeeting software, observers were able to visually see the responses being trained and discussed as each training set progressed. Observers were also allowed to hear the training through the software's audio function. In addition to observing the training of leadership virtually, representatives from Arizona also traveled to individual scoring sites to observe training in-person. This allowed Arizona to observe MI's training techniques and interact with project leadership. The State was able to provide additional guidance on scoring rationale during the training process. These observations allowed MI to further ensure reliability in the hand-scoring efforts.

Recruited staff followed established training methodologies to ensure the reliability and validity of scores. Scorers were trained as a group, not individually, and all scorers (whether experienced or not) were required to train on all the scoring sets and, at the end of training, pass the qualifying sets with acceptable scores to prove that they were able to understand and apply the criteria. Unless a scorer was trained and qualified for a project successfully, he or she was not permitted to score any student responses.

Each member of MI's scoring staff was required to qualify for the scoring of student responses based on standards established by Arizona following our vigorous training process. Each staff member was also expected to maintain a consistent level of scoring quality throughout the scoring effort or he or she was released from the project. MI continually monitored performance in order to guarantee scoring accuracy.

For math, MI trained scorers to hand-score a limited number of math items from the paper assessment that could not be machine-scored. Scoring leadership reviewed all hand-scored math items prior to training. Using the scoring rubrics provided from ITS, leadership provided feedback and questions to both AIR and Arizona to ensure consistency in training methodology. Math items were trained and scored individually with the use of the provided scoring rubrics. Qualified math scorers received training that included all possible answers to each individual item.

Math hand-scoring was monitored in the same way as essay scoring, with consistent read behind and validation sets incorporated into the daily scoring schedule to ensure that scorers were providing accurate scoring on a consistent basis.

10.2.4 MACHINE-SCORING VERIFICATION

In addition to the regular ELA hand scoring activities, MI also provided a percentage of second readings on items that were machine-scored. These read behind scores were used to help ensure consistency and reliability with the ELA machine-scoring. Responses requiring read behind were generated and sent to MI where the most experienced scorers, team leaders and scoring directors provided a second read verification. This process utilized blind scoring, with the scorer unaware of the first score provided by machine. Where scores from Reader 1 (machine) and Reader 2 (human) were in exact agreement or adjacent, the final item score was based on the initial machine read. Where scores from Reader 1 (machine) and Reader 2 (human) were not in exact agreement or adjacent, the final item score was based on the second human read.

11. QUALITY ASSURANCE PROCEDURES

Quality assurance procedures are enforced throughout all stages of AzMERIT test development, administration, and scoring and reporting of results. This section describes quality assurance procedures associated with

- Test construction
- Test production
- Answer document processing
- Data preparation
- Equating and scaling
- Scoring and reporting

Because quality assurance procedures pervade all aspects of test development, we note that discussion of quality assurance procedures is not limited to this section, but is also included in sections describing all phases of test development and implementation.

11.1 QUALITY ASSURANCE IN TEST CONSTRUCTION

Section 5.5 details the form construction process. Each form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test. The statistical targets ensure that students will receive scores of similar precision, regardless of which form of the test they receive.

The form construction process is managed through AIR's FormBuilder software which automates important form construction activities to ensure development of equated test forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers build a test map. This helps our content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

Immediately upon generation of a test form, the FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allows another opportunity to ensure that poorly performing items are not included in operational test forms.

When submitting test forms for review by ADE, AIR produces a form evaluation workbook that includes an evaluation summary checklist, as well as summary statistics and test characteristic graphs.

All bookmaps (test maps), key files, and conversion tables were produced directly from FormBuilder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents are generated directly from information maintained in ITS. The information stored in ITS is rigorously reviewed by multiple skilled reviewers, to protect against errors. Automated production of these critical files (such as key files) virtually eliminates opportunities for errors.

11.2 QUALITY ASSURANCE IN PAPER-DELIVERED TEST PRODUCTION

Camera-ready documents are prepared after the test items have been selected, composed in forms, and reviewed per the ADE's specifications.

Paper tests go through a traditional production process. The test booklet production process starts with the creation of test maps (also referred to as bookmaps). The test map is built in the Item Tracking System (ITS) and initiates the production of printed test forms. The process includes the following five steps:

1. The 1×1s (test items printed one per page) are generated based on the test map.
2. Blackline 1 is drafted and reviewed internally.
3. Blackline 1 is delivered to the Department for review and approval.
4. Should any changes be requested in the blackline 1 review, blackline 2s are produced, reviewed, and delivered to the Department.
5. The documents are brought to blueline (camera-ready copy).

Step 1 is entirely automated within ITS. ITS houses destination templates that define the format of the 1×1s and automatically generates these documents based on the test map. At this stage, items are proofread by internal editorial and test development staff and the Department. Additionally, they are reviewed to verify that all edits from previous rounds of review have been correctly implemented. Any changes required at this stage are entered directly into ITS to ensure consistency across all item uses.

Blackline 1 is a semi-automated process. With the appropriate destination template defined and 1×1 approval, ITS generates a Quark-readable document in the specified format. Through this integration, items are automatically styled with fonts, graphics, spacing, and other formatting specifications outlined in the Department's style guide. Our production staff may adjust page layout, including instructions, borders, and other elements, to meet the Department's guidelines. At this stage, reviewers check the document layout and formatting. Should any egregious errors be found in the content of an item, changes must be entered into ITS and the item must be re-exported to ensure consistent item use across all test forms. Changes to blackline 1 require a second blackline proof. Changes to subsequent blackline proofs require sign-off by senior management.

The final quality assurance step prior to printing is the blueline, or camera-ready copy, review stage. During this step, AIR and the Department's staff review proofs from the print vendor, verifying that the file to be printed matches the previously approved blackline proof. At AIR, in addition to reviews by test development and forms production staff, two members of the technical team—who have not seen the items previously—independently take the tests. This process forces a close look at the items and gives a final opportunity to verify the keys.

During the production and review process, test book blacklines are accompanied by answer document blacklines, which are produced by MI. Answer documents reflect the demographic fields required by the Department, as well as fields for pre-code labels and the scannable marks required for accurate data collection. The item sequence is based on test maps and corresponds directly with test books.

All blacklines in AIR's production queue are controlled by an electronic version-control server system that ensures that only the current version is immediately available to our production staff, preventing version-control errors. Like AIR's ITS, which controls and tracks all changes to items, this production system maintains historical records (including all older versions), which senior production staff can access if necessary. Each blackline after blackline 1 and the blueline (camera-ready copy) is automatically compared with the immediately preceding version using a PDF comparison tool that highlights all changes. This step has proved useful for identifying unintended changes made during the revision process. Such changes are difficult to detect because they can

appear anywhere in a document and may be subtle. The PDF comparison tool highlights these changes so differences between versions can be mapped to an intended revision. All materials delivered will go through this process, ensuring that the Department will receive error-free materials for review and that any changes requested by the Department are implemented promptly and accurately.

At each of the review stages, proofs will be accompanied by proof tickets that identify the document being reviewed, its review stage, the scheduled and actual delivery dates, and the return date. Sign-off by the Department will be required at each stage before proceeding with subsequent steps.

11.3 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

The production of computer-delivered assessments involves two distinct types of products, each of which follows an appropriate quality assurance process:

1. Content for online delivery shares some processes with paper versions, but also requires additional, unique steps.
2. Online test delivery software must deliver the content reliably (and, with the right tools, the accommodations, layouts, etc.).

11.3.1 PRODUCTION OF CONTENT

While the online workflow requires some additional steps, it actually removes a substantial amount of work from the time critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. A paper form must await final forms construction before blackline proofs can show how the item will look in the booklet. Online, the appearance of the item screen can be known with certainty before the final test form is ever constructed. This characteristic of online forms enables us to lock down the final presentation of each item well before forms are constructed. In turn, this moves the final blueprint review of items much earlier in the process, removing it from the critical path.

The production of computer-based tests includes five key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. Forms are finalized using the process described in Section 4.6, and final forms are approved in our FormBuilder software.
3. Complete test packages are created with our test packager, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
4. Forms are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the field.
5. The final system is deployed to a staging environment accessible to ADE for user acceptance testing and final review.

11.3.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The Item Tracking System (ITS) integrates directly with the test delivery system (TDS) display module, and displays each item exactly as it will appear to the student. This process is called web preview, and web preview is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the

student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent web preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will see it. Layout changes can be made after this process in two ways:

1. Content can be revised and re-approved for web display.
2. Online style sheets can change to revise the layout of all items on the test.

Both of these processes are subject to strict change control protocols to ensure that accidental changes are not introduced. Below, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the benefit of allowing final layout review much earlier in the process, reducing the work that must be done during the very busy period just before tests go live.

11.3.3 APPROVAL OF FINAL FORMS

Section 5.6 describes our process for constructing operational test forms, including the approval of test forms by ADE. The forms are built in FormBuilder (a component of our ITS), and upon approval, they are ready for preliminary publication.

11.3.4 PACKAGING

The test packaging system performs two simultaneous roles in the preparation of computer-based products: It compiles the form definitions and other information about how the test is to be administered (e.g., where any embedded field-test items might be inserted) and pulls together the content packaged during web approval.

The test packager assigns form identifiers to each form, evaluates the form against the blueprint, and performs a quality check against the content. The content quality check includes checks to see that every asset (e.g., graphics) referenced in the item is included in the package, confirms that the item has not changed since it was web approved, and ensures that the items have received all the approvals necessary for publication.

11.3.5 PLATFORM REVIEW

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

11.3.6 USER ACCEPTANCE TESTING AND FINAL REVIEW

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the test delivery system serves both a software evaluation and content approval role. The UAT period provides ADE with an opportunity to interact with the exact test with which the students will interact.

11.3.7 FUNCTIONALITY AND CONFIGURATION

The items, both in themselves and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of our system:

1. Testing on the system itself to ensure function, performance, and capacity
2. Capacity planning
3. Continuous monitoring

AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

11.4 QUALITY ASSURANCE IN DOCUMENT PROCESSING

11.4.1 SCANNING ACCURACY

When test documents were returned to be scored they must be first scanned. When they were scanned, a quality control sample of documents consisted of ten test cases per document type (normally between five and six hundred documents) were created so that all possible responses and all demographic grids were verified including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured method of scan testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner, editing process (validation and data correction), data transfer to the project database, and scoring were all accurate according to the reporting rules provided by ADE.

11.4.2 QUALITY ASSURANCE IN EDITING AND DATA INPUT

At a minimum, MI implemented, maintained, and constantly updated the following quality assurance controls:

- Score key verification Post analysis of item keys
- Response analyses to determine score frequency distribution by item verification of bank values of item statistics
- Live data checks to verify that data/results conform to approved specifications comprehensive software test plan

- Double data entry correction process to verify student response and demographic information report data verification
- Reviewed and proofread all electronic and printed report deliverables

MI utilized a double data correction process to achieve the highest level of quality and accuracy in both Arizona CBT and PBT assessment student data. Data correction operators used their sophisticated Data Inspection, Correction and Entry (DICE) application, which retrieved flagged data records and highlighted the problem field on a computer screen for resolution. The operator compared the highlighted data on the answer document template, retrieved the original document for resolution, and made any necessary correction.

After an operator corrected a flagged record, the same flagged record was routed to a second data correction operator who repeated the data correction process. After a flagged record was edited by two independent operators, the data correction application checked to verify that both operators made identical corrections. If the two corrections differed, the record was routed to a supervisor for a third and final resolution. Agreement rate statistics were generated for the individual data correction operators, allowing the supervisor to monitor their job performance. This process continued until all flagged records are examined and resolved.

Thorough training significantly improves the accuracy of data correction. To ensure that goal, MI trained their data correction staff on the use of the data correction application and on the specific validation errors and procedures associated with the specific project. Practice sets generated by the programming staff allowed data correction staff to learn on samples of answer documents that simulated the kinds of errors they were expected to correct for the actual assessment prior to actually processing live data. Additionally, each user had an electronic copy of the data correction user's guide for reference.

MI developed verification routines as part of their standard data validation to detect duplicate student tests in the assessment, whether in a single LEA (local educational agency) or across LEAs, and student moves between schools. MI staff then worked closely with ADE to resolve these discrepancies through processes called Barcode Processing and Tested Roster. These processes and the business rules governing them were described in a set of requirements developed in conjunction with ADE. They involved direct data transfer in several steps between the MI and ADE databases, with the goal of ensuring that each student final report was sent to the school where the test was taken, that it had accurate demographic data, and that the test reported was the correct test per the business rules.

11.5 QUALITY ASSURANCE IN DATA PREPARATION

AIR's test delivery system has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. For example, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the sorts of checks that are done for data review, but (a) they are done on operational data and (b) they are conducted in real time so that our psychometricians can catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM to the database of record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to ADE and their quality assurance contractor. AIR psychometricians ensure that data in the extract files matches the DoR prior to delivery to ADE.

11.6 QUALITY ASSURANCE IN TEST FORM EQUATING

Item information necessary for statistical and psychometric analyses is provided to ADE and HumRRO, ADE's independent quality assurance contractor, prior to test administration. Item information is published as part of the configuration of the online assessment system that AIR employs for administering, scoring, and reporting test scores. Information contained in these workbooks includes, but is not limited to, unique item ID used for item tracking, test form ID, location on the test form, correct answer, item difficulty, and information about the strand, standard, and benchmark each item measures. These item files are used in quality control checks of the assessment data scoring and analysis.

To ensure security, all data is shared using ADE's SFTP site.

Prior to operational work, AIR produces simulated datasets for the purpose of testing software and analysis procedures, and shares with ADE and the QA contractor. All parties complete a dry run of calibration and post-equating activities and compare results. The practice runs serve two functions:

1. To verify accuracy of program code and procedures.
2. To evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.

Following the completion of these activities and resolution of questions that arise, analysis specifications are finalized.

11.7 QUALITY ASSURANCE IN SCORING AND REPORTING

11.7.1 QUALITY ASSURANCE IN HAND SCORING

DOUBLE SCORING RATES, AGREEMENT RATES, VALIDITY SETS, AND ONGOING READ-BEHINDS.

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students.

MI's Virtual Scoring Center software (VSC), described in 10.2.1, provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer's performance every day to ensure that he or she is on target and conduct one-on-one retraining sessions when necessary. MI's quality assurance procedures allow scoring staff to identify struggling scorers very early and begin retraining immediately.

We monitor their scoring intensively to ensure all responses are scored accurately. If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, he or she is given interactive feedback and mentoring on the responses that have been scored incorrectly and is expected to change the scores. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

If a scorer's interrater agreement rate falls below the expected standard, the scorer will be re-trained. Should the scorer still be unable to score reliably, the scorer is assigned to another, non-Arizona-related project or dismissed.

In addition to using validity responses as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be culled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following Arizona's review and approval. MI periodically administers validity sets to each of MI's scorers working on the scoring effort. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the state.

With the VSC program, the way in which the student responses are presented prevents scorers from having any knowledge about which responses are being single or double read, or which responses are validity set responses. A performance threshold of 75% is set to specify validity agreement standards as well as the frequency and total number of validity responses evaluated by each scorer based on client specifications.

HANDSCORING QA MONITORING REPORTS

MI generates detailed scorer status reports for each scoring project utilizing a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Arizona. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available to Arizona 24 hours a day via a secure website. Project leadership reviews these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

Scorers are released when they are unable to demonstrate the ability to score responses according to the criteria and standards established by MI and Arizona and perform to the level of client expectation. Should Arizona request that certain responses be rescored, we are prepared to do so if necessary. The reporting system can produce a list of all the responses a selected scorer has scored. In these situations, all responses scored by a scorer during the time frame in question can be identified, reset, and released back into the scoring pool. The aberrant scorer's scores are deleted, and the responses are redistributed to other qualified scorers for rescoring.

MONITORING BY ARIZONA DEPARTMENT OF EDUCATION

ADE also directly observes MI activities, both on-site and virtually. MI provides virtual access to the training activities through the online training interface, as well as on-site training and on-site scoring. Arizona monitors the scoring process through the Client Command Center (CCC) with access to view and run specific reports during the scoring process. This ability to attend the training, qualification, and initial scoring virtually provides Arizona the most efficient use of oversight by reducing the travel requirements for on-site attendance for ADE staff.

IDENTIFYING, EVALUATING, AND INFORMING THE STATE ON ALERT PAPERS

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the examinee. We also flag potential security breaches identified during scoring. For possible dangerous

situations, scoring project management and staff employ a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

This process is also used to notify Arizona of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow up.

11.7.2 TEST SCORING

AIR verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each of these simulated students is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they provide a check of the full range of item responses and test scores in fixed-form tests as well. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the online reporting system, we merge item response data with the demographic information taken either from previous year assessment data, or if current year enrollment data is available by the time simulated data files are created, we can verify online reporting using current year testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the reporting system.

Specifications for generating simulated data files are included in the Analysis Specifications document submitted to and approved by ADE each year. Although ADE does not currently provide immediate reporting, review of all simulated data is scheduled to be completed prior to the opening of the test administration, so that the integrity of item administration, data capture, item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the test administration window, a series of Quality Assurance Reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. In the context of adaptive test administrations, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to specifications.

An additional set of forensic analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school level that may indicate cheating. The quality assurance reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the test window to ensure that test administrations conform to blueprint and items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit 11.6.2.1 presents an overview of the quality assurance (QA) reports.

Exhibit 11.6.2.1 Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Analysis Report	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Forensic Analysis	To monitor testing irregularities	Early detection of testing irregularities

ITEM ANALYSIS REPORT

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For multiple-choice items, the proportion of students selecting each response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for selected-response items and polyserial correlations for polytomous constructed response, performance, and technology items are computed. AIR psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between observed and expected score given the student's ability is computed for each item. The residuals for each are averaged across all students, and the average residual is used to flag an item. The item fit statistic is computed as follows:

Let X_{ij} be the variable for the response of student j to item i , and $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ be the probability that student j gets a score of x_{ij} to item i given his or her ability estimate $\hat{\theta}_j$. $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ is calculated using three parameter logistic IRT model

$$P(X_{ij} = x_{ij}|\hat{\theta}_j) = c_i + (-c_i) \frac{\exp(a_i(\hat{\theta}_j - b_i))}{1 + \exp(a_i(\hat{\theta}_j - b_i))},$$

where a_i , b_i , and c_i are parameters of item i . If item i is a polytomously scored item, $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ is calculated using the Generalized Partial Credit model,

$$P(X_{ij} = x_{ij} | \hat{\theta}_j) = \frac{\exp \sum_{k=0}^{x_{ij}} a_i(\hat{\theta}_j - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l a_i(\hat{\theta}_j - b_{ki})}$$

The expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as

$$E(X_{ij} | \hat{\theta}_j) = \sum_{x_{ij}=0}^{m_i} x_{ij} P(X_{ij} = x_{ij} | \hat{\theta}_j).$$

For item i , the residual between observed and expected score for student j is defined as

$$\delta_{ij} = x_{ij} - E(X_{ij} | \hat{\theta}_j).$$

The statistic δ_{ij} is aggregated across all n students for item i ,

$$\bar{\delta}_i = \frac{1}{n} \sum_{i=1}^n (\delta_{ij}).$$

The report can be configured to report all items or flag and report only those items where the fit index is above a given threshold (e.g., items could be flagged when

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > .96$$

where $se(\bar{\delta}_j) = \frac{SD(\delta_{ij})}{\sqrt{n}}$.

FORENSIC ANALYSIS

Another component in the suite of QA reports is geared toward detecting testing irregularities that may indicate possible cheating. The forensic analysis components of the QA reports are described in detail in Section 6.6. Evidence evaluated includes changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

11.7.3 REPORTING

Scores for online assessments are assigned by automated systems in real time. For machine scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field-testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The hand-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Once both online and hand scoring items have passed through their validity and quality checks, the

hand-scored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the quality monitor (QM) system where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating performance-level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively prior to deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

After passing through the series of validation checks in the QM system, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring there is only one place where the “official” record is stored. Only after scores have passed the QM checks and are uploaded to the DoR are they passed to the Online Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Online Reporting System until it passes all of the QM system’s validation checks and ADE’s independent data verification checks.

12. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Council of Chief State School Officers. (2014). *Criteria for Procuring and Evaluating High-Quality Assessments*. Retrieved from www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf.
- Dragow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(), 67–86.
- Estrada, S., Burnham, C., Feld, J. K., Bergan, J. R., & Bergan, J. R. (2015). *Research Commentary: Can Local Assessment Data be Successfully Used as Part of an Arizona A-F Accountability System?* Assessment Technology Incorporated.
- Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, 21(3), 187-206.
- Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, 4, 231-246.
- Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Linacre, J. M. (2004). *A user's guide to WINSTEPS: Rasch-Model Computer Program*. Chicago, IL: MESA Press.
- Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 16(4), 247-260.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179-197.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52(2), 443-451.
- McLaughlin, D., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005). *Using State Assessments to Impute Achievement of Students Absent from NAEP: An Empirical Study in Four States*. Washington, DC: American Institutes for Research. Retrieved from www.air.org/files/McLaughlin_AbsentStudents.pdf.

- Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, 1, 115-135.
- Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2-3), 170-187.
- Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Phillipine Statistician*, 52(4), 81-92.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66(3), 331-342.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13(4), 265-276.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
- Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(), -10.
- Scott, L. (2015). Analysis of Mode Comparability of AzMERIT's Online and Paper Administrations for Spring 2015. In Arizona Department of Education, *Recommending AzMERIT Performance Standards* (pp. I-28 – I-40), Retrieved from <https://cms.azed.gov/home/GetDocumentFile?id=5846d5b4aadebe0cf0337f5e>.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. In annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Webb, N. L. (2002). *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States*. Washington, DC: Council of Chief State School Officers.
- Webb, N. (2005, April). Issues related to judging the alignment of curriculum standards and assessments. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27(7), 909-921.



English Language Arts Assessment Blueprint

Grade 3		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 4		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 5		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 6		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 7		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 8		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 9		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 10		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 11		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Listening Standards will only be assessed on the computer-based assessment.

In Grades 3-5 some items in the Reading and Language Strands will also be aligned to the standards for Reading: Foundational Skills.

Percentage of Points by Depth of Knowledge Level

Grade	DOK Level 1	DOK Level 2	DOK Level 3	DOK Level 4
3-11	10%-20%	50%-60%	15%-25%	16%-19% (Writing)

For more information go to www.azed.gov/AzMERIT

Grade 3		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	49%	53%
Number and Operations-Fractions	18%	22%
Measurement, Data, and Geometry	26%	30%

Grade 6		
Domain	Min.	Max.
Ratio and Proportional Relationships	19%	23%
The Number System	25%	29%
Expressions and Equations	29%	33%
Geometry, Statistics and Probability	17%	21%

Algebra I		
Conceptual Categories	Min.	Max.
Algebra	40%	44%
Functions	36%	40%
Statistics	17%	21%

Percentage of Points by Depth of Knowledge Level			
Grade	DOK Level 1	DOK Level 2	DOK Level 3
3-11	10%-20%	60%-70%	12%-30%

Grade 4		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	46%	54%
Number and Operations-Fractions	29%	33%
Measurement, Data, and Geometry	15%	19%

Grade 7		
Domain	Min.	Max.
Ratio and Proportional Relationships	19%	23%
The Number System	19%	23%
Expressions and Equations	23%	27%
Geometry, Statistics and Probability	27%	35%

Geometry		
Domain	Min.	Max.
Congruence	23%	27%
Similarity, Right Triangles and Trigonometry	27%	31%
Circles , Geometric Measurement and Geometric Properties with Equations	23%	27%
Modeling with Geometry	17%	21%

Within a test, approximately 70% of the assessment will be on major content within that grade or course.

Revised by ADE on 8/19/15

Grade 5		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	38%	42%
Number and Operations-Fractions	31%	35%
Measurement, Data, and Geometry	24%	28%

Grade 8		
Domain	Min.	Max.
Expressions and Equations	32%	36%
Functions	21%	25%
Geometry	23%	27%
Statistics and Probability and The Number System	15%	19%

Algebra II		
Conceptual Categories	Min.	Max.
Algebra	34%	38%
Functions	32%	36%
Statistics	27%	31%

For more information go to www.azed.gov/AzMERIT

Appendix B.1a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	13733.772	1804				
Metric	13956.001	1847	Configural	222.230 (43)	< .01	.000
Scalar	14638.886	1890	Metric	682.885 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7117.742	1804				
Metric	7271.625	1847	Configural	153.883 (43)	< .01	.000
Scalar	7347.689	1890	Metric	76.064 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	11819.287	1804				
Metric	12378.423	1847	Configural	559.136 (43)	< .01	.000
Scalar	12648.918	1890	Metric	270.495 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6852.658	1804				
Metric	6900.130	1847	Configural	47.473 (43)	0.30	.000
Scalar	7101.063	1890	Metric	200.932 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7034.814	1804				
Metric	7278.311	1847	Configural	243.498 (43)	< .01	.000
Scalar	7436.232	1890	Metric	157.921 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7047.919	1804				
Metric	7113.779	1847	Configural	65.860 (43)	0.01	.001
Scalar	7172.388	1890	Metric	58.609 (43)	0.06	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13509.610	1804				
Metric	13867.551	1847	Configural	357.941 (43)	< .01	.000
Scalar	14304.595	1890	Metric	437.045 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	13726.756	1804				
Metric	14038.892	1847	Configural	312.136 (43)	< .01	.000
Scalar	14103.428	1890	Metric	64.536 (43)	0.02	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	13549.468	1804				
Metric	14031.524	1847	Configural	482.057 (43)	< .01	.000
Scalar	14284.069	1890	Metric	252.545 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13220.548	1804				
Metric	14268.173	1847	Configural	1047.625 (43)	< .01	.001
Scalar	14879.019	1890	Metric	610.846 (43)	< .01	.001

Appendix B.1b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	14638.886	1890	< .01	0.945	0.039
Model B-1	7347.689	1890	< .01	0.954	0.035
Model B-2	12648.918	1890	< .01	0.941	0.039
Model B-3	7101.063	1890	< .01	0.956	0.033
Model B-4	7436.232	1890	< .01	0.971	0.023
Model B-5	7172.388	1890	< .01	0.954	0.034
Model C	14304.595	1890	< .01	0.944	0.036
Model D	14103.428	1890	< .01	0.946	0.039
Model E	14284.069	1890	< .01	0.966	0.025
Model F	14879.019	1890	< .01	0.956	0.027

Appendix B.2a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	10163.750	1804				
Metric	10328.986	1847	Configural	165.236 (43)	< .01	.000
Scalar	10978.449	1890	Metric	649.463 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5563.647	1804				
Metric	5758.299	1847	Configural	194.652 (43)	< .01	.000
Scalar	5871.122	1890	Metric	112.823 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8716.399	1804				
Metric	9243.675	1847	Configural	527.276 (43)	< .01	.001
Scalar	9913.604	1890	Metric	669.929 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5405.843	1804				
Metric	5488.063	1847	Configural	82.220 (43)	< .01	.001
Scalar	5701.817	1890	Metric	213.754 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5509.142	1804				
Metric	5806.752	1847	Configural	297.610 (43)	< .01	.001
Scalar	6041.940	1890	Metric	235.189 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5468.639	1804				
Metric	5521.446	1847	Configural	52.807	0.15	.001
Scalar	5556.291	1890	Metric	34.845	0.81	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10234.397	1804				
Metric	10695.816	1847	Configural	461.419 (43)	< .01	.000
Scalar	11194.600	1890	Metric	498.784 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	10133.666	1804				
Metric	10554.761	1847	Configural	421.096 (43)	< .01	.001
Scalar	10709.722	1890	Metric	154.961 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	10034.335	1804				
Metric	10648.370	1847	Configural	614.035 (43)	< .01	.001
Scalar	10960.887	1890	Metric	312.516 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9881.793	1804				
Metric	11189.067	1847	Configural	1307.274 (43)	< .01	.002
Scalar	11780.373	1890	Metric	591.305 (43)	< .01	.000

Appendix B.2b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10978.449	1890	< .01	0.947	0.036
Model B-1	5871.122	1890	< .01	0.986	0.016
Model B-2	9913.604	1890	< .01	0.968	0.024
Model B-3	5701.817	1890	< .01	0.947	0.031
Model B-4	6041.940	1890	< .01	0.984	0.016
Model B-5	5556.291	1890	< .01	0.974	0.021
Model C	11194.600	1890	< .01	0.943	0.033
Model D	10709.722	1890	< .01	0.946	0.036
Model E	10960.887	1890	< .01	0.986	0.016
Model F	11780.373	1890	< .01	0.938	0.034

Appendix B.3a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	11011.824	1804				
Metric	11305.868	1847	Configural	294.044 (43)	< .01	.000
Scalar	11780.179	1890	Metric	474.311 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6128.336	1804				
Metric	6312.812	1847	Configural	184.476 (43)	< .01	.000
Scalar	6406.820	1890	Metric	94.008 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9646.302	1804				
Metric	10044.148	1847	Configural	397.846 (43)	< .01	.001
Scalar	10327.027	1890	Metric	282.879 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6076.608	1804				
Metric	6143.701	1847	Configural	67.094 (43)	0.01	.000
Scalar	6278.654	1890	Metric	134.953 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6160.021	1804				
Metric	6479.098	1847	Configural	319.077 (43)	< .01	.001
Scalar	6674.439	1890	Metric	195.341 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6003.154	1804				
Metric	6075.136	1847	Configural	71.982 (43)	< .01	.000
Scalar	6129.819	1890	Metric	54.682 (43)	0.11	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10971.222	1804				
Metric	11588.948	1847	Configural	617.727 (43)	< .01	.000
Scalar	12189.102	1890	Metric	600.153 (43)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11119.735	1804				
Metric	11347.250	1847	Configural	227.515 (43)	< .01	.000
Scalar	11426.483	1890	Metric	79.233 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11049.637	1804				
Metric	11494.150	1847	Configural	444.512 (43)	< .01	.000
Scalar	11680.610	1890	Metric	186.460 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	10695.809	1804				
Metric	12027.699	1847	Configural	1331.889 (43)	< .01	.002
Scalar	12617.063	1890	Metric	589.364 (43)	< .01	.001

Appendix B.3b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	11780.179	1890	< .01	0.970	0.031
Model B-1	6406.820	1890	< .01	0.986	0.016
Model B-2	10327.027	1890	< .01	0.964	0.032
Model B-3	6278.654	1890	< .01	0.969	0.027
Model B-4	6674.439	1890	< .01	0.984	0.016
Model B-5	6129.819	1890	< .01	0.970	0.027
Model C	12189.102	1890	< .01	0.966	0.029
Model D	11426.483	1890	< .01	0.970	0.031
Model E	11680.610	1890	< .01	0.987	0.015
Model F	12617.063	1890	< .01	0.981	0.017

Appendix B.4a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	9698.728	1804				
Metric	10093.239	1847	Configural	394.510 (43)	< .01	.001
Scalar	10841.059	1890	Metric	747.820 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5611.090	1804				
Metric	5743.655	1847	Configural	132.565 (43)	< .01	.000
Scalar	5860.740	1890	Metric	117.085 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8409.068	1804				
Metric	8938.388	1847	Configural	529.319 (43)	< .01	.001
Scalar	9380.459	1890	Metric	442.072 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5466.053	1804				
Metric	5516.666	1847	Configural	50.614	0.20	.000
Scalar	5655.420	1890	Metric	138.753	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5652.413	1804				
Metric	5960.932	1847	Configural	308.519 (43)	< .01	.000
Scalar	6181.560	1890	Metric	220.628 (43)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5444.331	1804				
Metric	5480.300	1847	Configural	35.969 (43)	0.77	.001
Scalar	5526.887	1890	Metric	46.587 (43)	0.33	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9547.579	1804				
Metric	10051.098	1847	Configural	503.520 (43)	< .01	.001
Scalar	10564.126	1890	Metric	513.028 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9673.726	1804				
Metric	10094.442	1847	Configural	420.715 (43)	< .01	.001
Scalar	10208.083	1890	Metric	113.641 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9627.312	1804				
Metric	9925.714	1847	Configural	298.403 (43)	< .01	.000
Scalar	10293.927	1890	Metric	368.213 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9313.958	1804				
Metric	9980.107	1847	Configural	666.149 (43)	< .01	.000
Scalar	10761.774	1890	Metric	781.667 (43)	< .01	.001

Appendix B.4b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10841.059	1890	< .01	0.880	0.032
Model B-1	5860.740	1890	< .01	0.966	0.030
Model B-2	9380.459	1890	< .01	0.961	0.033
Model B-3	5655.420	1890	< .01	0.989	0.013
Model B-4	6181.560	1890	< .01	0.986	0.016
Model B-5	5526.887	1890	< .01	0.969	0.027
Model C	10564.126	1890	< .01	0.870	0.031
Model D	10208.083	1890	< .01	0.890	0.030
Model E	10293.927	1890	< .01	0.880	0.031
Model F	10761.774	1890	< .01	0.986	0.015

Appendix B.5a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	7386.057	1804				
Metric	7573.876	1847	Configural	187.819 (43)	< .01	.000
Scalar	8502.762	1890	Metric	928.886 (43)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	4703.545	1804				
Metric	4856.004	1847	Configural	152.459 (43)	< .01	.000
Scalar	4973.729	1890	Metric	117.725 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6684.172	1804				
Metric	7103.969	1847	Configural	419.798 (43)	< .01	.000
Scalar	7451.795	1890	Metric	347.826 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4718.096	1804				
Metric	4764.669	1847	Configural	46.573 (43)	0.33	.000
Scalar	4939.561	1890	Metric	174.891 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	4623.233	1804				
Metric	4784.118	1847	Configural	160.885 (43)	< .01	.000
Scalar	4959.031	1890	Metric	174.913 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	4665.775	1804				
Metric	4707.603	1847	Configural	41.828 (43)	0.52	.001
Scalar	4769.442	1890	Metric	61.839 (43)	0.03	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7391.785	1804				
Metric	7699.688	1847	Configural	307.903 (43)	< .01	.001
Scalar	8152.133	1890	Metric	452.445 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7522.564	1804				
Metric	7856.478	1847	Configural	333.913 (43)	< .01	.000
Scalar	7923.163	1890	Metric	66.685 (43)	0.01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7541.658	1804				
Metric	7791.811	1847	Configural	250.154 (43)	< .01	.000
Scalar	8127.541	1890	Metric	335.730 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7431.922	1804				
Metric	7840.706	1847	Configural	408.784 (43)	< .01	.001
Scalar	8519.549	1890	Metric	678.843 (43)	< .01	.001

Appendix B.5b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	8502.762	1890	< .01	0.967	0.030
Model B-1	4973.729	1890	< .01	0.988	0.013
Model B-2	7451.795	1890	< .01	0.961	0.029
Model B-3	4939.561	1890	< .01	0.966	0.024
Model B-4	4959.031	1890	< .01	0.988	0.013
Model B-5	4769.442	1890	< .01	0.969	0.023
Model C	8152.133	1890	< .01	0.989	0.013
Model D	7923.163	1890	< .01	0.966	0.028
Model E	8127.541	1890	< .01	0.990	0.012
Model F	8519.549	1890	< .01	0.988	0.013

Appendix B.6a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	14027.579	1804				
Metric	14325.678	1847	Configural	298.099 (43)	< .01	.000
Scalar	15084.733	1890	Metric	759.055 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7738.802	1804				
Metric	7889.631	1847	Configural	150.829 (43)	< .01	.000
Scalar	8035.171	1890	Metric	145.540 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	12202.711	1804				
Metric	12716.273	1847	Configural	513.562 (43)	< .01	.000
Scalar	12922.089	1890	Metric	205.816 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	7613.130	1804				
Metric	7667.109	1847	Configural	53.979 (43)	0.12	.000
Scalar	7793.346	1890	Metric	126.237 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7776.336	1804				
Metric	7970.933	1847	Configural	194.597 (43)	< .01	.000
Scalar	8130.334	1890	Metric	159.402 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7529.398	1804				
Metric	7562.979	1847	Configural	33.581 (43)	0.85	.001
Scalar	7616.296	1890	Metric	53.317 (43)	0.13	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13905.703	1804				
Metric	14240.091	1847	Configural	334.388 (43)	< .01	.000
Scalar	14864.115	1890	Metric	624.024 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	14078.981	1804				
Metric	14327.866	1847	Configural	248.885 (43)	< .01	.000
Scalar	14421.238	1890	Metric	93.372 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	14095.956	1804				
Metric	14211.952	1847	Configural	115.996 (43)	< .01	.000
Scalar	14414.906	1890	Metric	202.954 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13861.740	1804				
Metric	14406.544	1847	Configural	544.805 (43)	< .01	.000
Scalar	15104.898	1890	Metric	698.353 (43)	< .01	.000

Appendix B.6b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	15084.733	1890	< .01	0.948	0.044
Model B-1	8035.171	1890	< .01	0.957	0.041
Model B-2	12922.089	1890	< .01	0.951	0.044
Model B-3	7793.346	1890	< .01	0.957	0.037
Model B-4	8130.334	1890	< .01	0.958	0.040
Model B-5	7616.296	1890	< .01	0.961	0.036
Model C	14864.115	1890	< .01	0.947	0.038
Model D	14421.238	1890	< .01	0.947	0.042
Model E	14414.906	1890	< .01	0.950	0.035
Model F	15104.898	1890	< .01	0.971	0.038

Appendix B.7a. Global Model Fit Indices of Measurement Invariance Tests for Grade 9 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	9785.018	1978				
Metric	10049.839	2023	Configural	264.821 (45)	< .01	.000
Scalar	10769.979	2068	Metric	720.140 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5881.394	1978				
Metric	5997.772	2023	Configural	116.378 (45)	< .01	.000
Scalar	6098.463	2068	Metric	100.691 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8602.375	1978				
Metric	9013.777	2023	Configural	411.402 (45)	< .01	.000
Scalar	9358.174	2068	Metric	344.397 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5811.836	1978				
Metric	5858.039	2023	Configural	46.203 (45)	0.42	.001
Scalar	5961.451	2068	Metric	103.412 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5876.617	1978				
Metric	6018.198	2023	Configural	141.581 (45)	< .01	.000
Scalar	6266.719	2068	Metric	248.521 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5702.498	1978				
Metric	5741.660	2023	Configural	39.162 (45)	0.72	.001
Scalar	5776.327	2068	Metric	34.667 (45)	0.87	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9819.145	1978				
Metric	10020.292	2023	Configural	201.147 (45)	< .01	.000
Scalar	10344.222	2068	Metric	323.930 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9860.151	1978				
Metric	10147.583	2023	Configural	287.432 (45)	< .01	.000
Scalar	10258.257	2068	Metric	110.674 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9970.516	1978				
Metric	10173.098	2023	Configural	202.582 (45)	< .01	.000
Scalar	10314.595	2068	Metric	141.497 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9912.833	1978				
Metric	10125.393	2023	Configural	212.560 (45)	< .01	.000
Scalar	10352.306	2068	Metric	226.913 (45)	< .01	.000

Appendix B.7b. Global Model Fit Indices of Scalar Invariance Model for Grade 9 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10769.979	2068	< .01	0.965	0.023
Model B-1	6098.463	2068	< .01	0.972	0.019
Model B-2	9358.174	2068	< .01	0.963	0.023
Model B-3	5961.451	2068	< .01	0.974	0.017
Model B-4	6266.719	2068	< .01	0.973	0.019
Model B-5	5776.327	2068	< .01	0.978	0.015
Model C	10344.222	2068	< .01	0.970	0.019
Model D	10258.257	2068	< .01	0.967	0.022
Model E	10314.595	2068	< .01	0.973	0.018
Model F	10352.306	2068	< .01	0.986	0.012

Appendix B.8a. Global Model Fit Indices of Measurement Invariance Tests for Grade 10 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	11398.967	1978				
Metric	11643.610	2023	Configural	244.644 (45)	< .01	.000
Scalar	12454.369	2068	Metric	810.759 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6787.701	1978				
Metric	6887.953	2023	Configural	100.253 (45)	< .01	.000
Scalar	7000.945	2068	Metric	112.992 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9962.149	1978				
Metric	10396.724	2023	Configural	434.575 (45)	< .01	.001
Scalar	10754.618	2068	Metric	357.894 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6718.523	1978				
Metric	6794.553	2023	Configural	76.030 (45)	< .01	.000
Scalar	6955.160	2068	Metric	160.607 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6677.657	1978				
Metric	6909.453	2023	Configural	231.797 (45)	< .01	.001
Scalar	7122.427	2068	Metric	212.974 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6731.650	1978				
Metric	6784.798	2023	Configural	53.148 (45)	0.19	.000
Scalar	6837.360	2068	Metric	52.562 (45)	0.20	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	11516.948	1978				
Metric	11743.401	2023	Configural	226.452 (45)	< .01	.000
Scalar	12092.078	2068	Metric	348.678 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11482.165	1978				
Metric	11785.845	2023	Configural	303.680 (45)	< .01	.000
Scalar	11878.384	2068	Metric	92.539 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11538.971	1978				
Metric	11651.848	2023	Configural	112.877 (45)	< .01	.000
Scalar	11768.221	2068	Metric	116.373 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	11479.163	1978				
Metric	11665.006	2023	Configural	185.844 (45)	< .01	.000
Scalar	11828.482	2068	Metric	163.476 (45)	< .01	.000

Appendix B.8b. Global Model Fit Indices of Scalar Invariance Model for Grade 10 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	12454.369	2068	< .01	0.962	0.036
Model B-1	7000.945	2068	< .01	0.960	0.021
Model B-2	10754.618	2068	< .01	0.947	0.025
Model B-3	6955.160	2068	< .01	0.958	0.021
Model B-4	7122.427	2068	< .01	0.960	0.021
Model B-5	6837.360	2068	< .01	0.963	0.018
Model C	12092.078	2068	< .01	0.965	0.031
Model D	11878.384	2068	< .01	0.965	0.035
Model E	11768.221	2068	< .01	0.978	0.028
Model F	11828.482	2068	< .01	0.978	0.027

Appendix B.9a. Global Model Fit Indices of Measurement Invariance Tests for Grade 11 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	7351.566	1978				
Metric	7591.812	2023	Configural	240.246 (45)	< .01	.000
Scalar	8303.823	2068	Metric	712.011 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5120.192	1978				
Metric	5209.060	2023	Configural	88.868 (45)	< .01	.000
Scalar	5319.957	2068	Metric	110.897 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6619.298	1978				
Metric	6981.561	2023	Configural	362.263 (45)	< .01	.001
Scalar	7298.593	2068	Metric	317.032 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4973.414	1978				
Metric	5009.619	2023	Configural	36.205 (45)	0.82	.000
Scalar	5129.101	2068	Metric	119.482 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5244.742	1978				
Metric	5422.163	2023	Configural	177.422 (45)	< .01	.001
Scalar	5688.003	2068	Metric	265.839 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5065.796	1978				
Metric	5119.452	2023	Configural	53.657 (45)	0.18	.000
Scalar	5145.823	2068	Metric	26.371 (45)	0.99	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7513.769	1978				
Metric	7678.313	2023	Configural	164.544 (45)	< .01	.000
Scalar	8058.493	2068	Metric	380.180 (45)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7488.459	1978				
Metric	7754.618	2023	Configural	266.159 (45)	< .01	.001
Scalar	7914.265	2068	Metric	159.647 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7717.892	1978				
Metric	7812.260	2023	Configural	94.368 (45)	< .01	.000
Scalar	7923.082	2068	Metric	110.822 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7704.274	1978				
Metric	7847.220	2023	Configural	142.946 (45)	< .01	.000
Scalar	7993.301	2068	Metric	146.081 (45)	< .01	.000

Appendix B.9b. Global Model Fit Indices of Scalar Invariance Model for Grade 11 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	8303.823	2068	< .01	0.959	0.034
Model B-1	5319.957	2068	< .01	0.973	0.020
Model B-2	7298.593	2068	< .01	0.964	0.023
Model B-3	5129.101	2068	< .01	0.975	0.018
Model B-4	5688.003	2068	< .01	0.972	0.021
Model B-5	5145.823	2068	< .01	0.977	0.016
Model C	8058.493	2068	< .01	0.969	0.027
Model D	7914.265	2068	< .01	0.960	0.032
Model E	7923.082	2068	< .01	0.985	0.014
Model F	7993.301	2068	< .01	0.983	0.015

Appendix B.10a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	178020.127	1890				
Metric	178975.724	1934	Configural	955.597 (44)	< .01	.001
Scalar	183551.597	1978	Metric	4575.874 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	74981.430	1890				
Metric	76809.764	1934	Configural	1828.334 (44)	< .01	.000
Scalar	77843.804	1978	Metric	1034.040 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	143550.580	1890				
Metric	149168.011	1934	Configural	5617.431 (44)	< .01	.000
Scalar	154151.396	1978	Metric	4983.385 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	71303.946	1890				
Metric	71673.151	1934	Configural	369.205 (44)	< .01	.001
Scalar	72140.060	1978	Metric	466.909 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	74961.905	1890				
Metric	77173.889	1934	Configural	2211.984 (44)	< .01	.000
Scalar	78790.231	1978	Metric	1616.342 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	71635.066	1890				
Metric	71792.403	1934	Configural	157.337 (44)	< .01	.000
Scalar	71909.228	1978	Metric	116.825 (44)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	170307.577	1890				
Metric	177157.840	1934	Configural	6850.263 (44)	< .01	.001
Scalar	180389.601	1978	Metric	3231.762 (44)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	175711.661	1890				
Metric	179583.439	1934	Configural	3871.778 (44)	< .01	.000
Scalar	181201.055	1978	Metric	1617.616 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	174317.255	1890				
Metric	180824.968	1934	Configural	6507.712 (44)	< .01	.000
Scalar	182097.212	1978	Metric	1272.244 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	165211.122	1890				
Metric	181890.732	1934	Configural	16679.610 (44)	< .01	.002
Scalar	184706.069	1978	Metric	2815.336 (44)	< .01	.001

Appendix B.10b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	183551.597	1978	< .01	0.917	0.058
Model B-1	77843.804	1978	< .01	0.913	0.056
Model B-2	154151.396	1978	< .01	0.909	0.056
Model B-3	72140.060	1978	< .01	0.922	0.050
Model B-4	78790.231	1978	< .01	0.917	0.055
Model B-5	71909.228	1978	< .01	0.915	0.054
Model C	180389.601	1978	< .01	0.905	0.058
Model D	181201.055	1978	< .01	0.914	0.057
Model E	182097.212	1978	< .01	0.916	0.057
Model F	184706.069	1978	< .01	0.906	0.057

Appendix B.11a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	78513.118	1890				
Metric	79737.405	1934	Configural	1224.287 (44)	< .01	.000
Scalar	84155.308	1978	Metric	4417.903 (44)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	31709.452	1890				
Metric	33546.915	1934	Configural	1837.464 (44)	< .01	.000
Scalar	34341.647	1978	Metric	794.731 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	62665.155	1890				
Metric	66540.482	1934	Configural	3875.328 (44)	< .01	.001
Scalar	69648.954	1978	Metric	3108.472 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	29285.910	1890				
Metric	29504.006	1934	Configural	218.096 (44)	< .01	.000
Scalar	30032.709	1978	Metric	528.703 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	31111.359	1890				
Metric	32984.692	1934	Configural	1873.333 (44)	< .01	.001
Scalar	34001.235	1978	Metric	1016.544 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	29303.472	1890				
Metric	29364.693	1934	Configural	61.220 (44)	.044	.000
Scalar	29494.527	1978	Metric	129.835 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	74429.878	1890				
Metric	80370.099	1934	Configural	5940.221 (44)	< .01	.001
Scalar	84016.074	1978	Metric	3645.976 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	77747.392	1890				
Metric	80662.164	1934	Configural	2914.772 (44)	< .01	.000
Scalar	81460.601	1978	Metric	798.437 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	76537.210	1890				
Metric	82372.857	1934	Configural	5835.648 (44)	< .01	.001
Scalar	83624.630	1978	Metric	1251.772 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	71418.760	1890				
Metric	84391.883	1934	Configural	12973.123 (44)	< .01	.002
Scalar	87919.282	1978	Metric	3527.398 (44)	< .01	.001

Appendix B.11b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	84155.308	1978	< .01	0.962	0.035
Model B-1	34341.647	1978	< .01	0.961	0.032
Model B-2	69648.954	1978	< .01	0.957	0.033
Model B-3	30032.709	1978	< .01	0.960	0.029
Model B-4	34001.235	1978	< .01	0.961	0.031
Model B-5	29494.527	1978	< .01	0.964	0.029
Model C	84016.074	1978	< .01	0.959	0.033
Model D	81460.601	1978	< .01	0.962	0.034
Model E	83624.630	1978	< .01	0.964	0.033
Model F	87919.282	1978	< .01	0.957	0.033

Appendix B.12a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	72135.653	1890				
Metric	73067.921	1934	Configural	932.268 (44)	< .01	.000
Scalar	76282.010	1978	Metric	3214.089 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	29168.642	1890				
Metric	31046.373	1934	Configural	1877.730 (44)	< .01	.001
Scalar	31621.282	1978	Metric	574.909 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	57946.466	1890				
Metric	62395.316	1934	Configural	4448.850 (44)	< .01	.001
Scalar	64315.473	1978	Metric	1920.157 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	26768.137	1890				
Metric	27112.072	1934	Configural	343.935 (44)	< .01	.000
Scalar	27481.486	1978	Metric	369.413 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	29519.527	1890				
Metric	32104.916	1934	Configural	2585.389 (44)	< .01	.001
Scalar	32872.831	1978	Metric	767.915 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26763.882	1890				
Metric	26846.021	1934	Configural	82.140 (44)	< .01	.001
Scalar	26943.191	1978	Metric	97.169 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	68366.428	1890				
Metric	74117.853	1934	Configural	5751.424 (44)	< .01	.001
Scalar	79295.211	1978	Metric	5177.358 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	71225.495	1890				
Metric	74139.995	1934	Configural	2914.500 (44)	< .01	.000
Scalar	74628.714	1978	Metric	488.719 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	70880.121	1890				
Metric	74815.092	1934	Configural	3934.970 (44)	< .01	.000
Scalar	77058.496	1978	Metric	2243.404 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	66416.026	1890				
Metric	77274.731	1934	Configural	10858.705 (44)	< .01	.002
Scalar	83493.323	1978	Metric	6218.592 (44)	< .01	.001

Appendix B.12b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	76282.010	1978	< .01	0.968	0.032
Model B-1	31621.282	1978	< .01	0.968	0.029
Model B-2	64315.473	1978	< .01	0.964	0.031
Model B-3	27481.486	1978	< .01	0.968	0.027
Model B-4	32872.831	1978	< .01	0.968	0.029
Model B-5	26943.191	1978	< .01	0.970	0.027
Model C	79295.211	1978	< .01	0.963	0.031
Model D	74628.714	1978	< .01	0.967	0.031
Model E	77058.496	1978	< .01	0.968	0.030
Model F	83493.323	1978	< .01	0.962	0.031

Appendix B.13a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	90030.261	2068				
Metric	91811.680	2114	Configural	1781.419 (46)	< .01	.000
Scalar	98752.423	2160	Metric	6940.743 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	37453.505	2068				
Metric	40103.361	2114	Configural	2649.856 (46)	< .01	.001
Scalar	40679.803	2160	Metric	576.442 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	71421.466	2068				
Metric	78474.546	2114	Configural	7053.080 (46)	< .01	.002
Scalar	80280.423	2160	Metric	1805.877 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	35345.538	2068				
Metric	35560.721	2114	Configural	215.182 (46)	< .01	.001
Scalar	35866.572	2160	Metric	305.851 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	37208.036	2068				
Metric	39896.741	2114	Configural	2688.705 (46)	< .01	.001
Scalar	40737.236	2160	Metric	840.495 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	34775.418	2068				
Metric	34927.089	2114	Configural	151.671 (46)	< .01	.000
Scalar	34986.399	2160	Metric	59.310 (46)	.090	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	80371.915	2068				
Metric	87601.812	2114	Configural	7229.897 (46)	< .01	.001
Scalar	96412.066	2160	Metric	8810.254 (46)	< .01	.002
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	88516.533	2068				
Metric	93104.084	2114	Configural	4587.551 (46)	< .01	.001
Scalar	93599.712	2160	Metric	495.628 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	86690.679	2068				
Metric	91830.076	2114	Configural	5139.397 (46)	< .01	.000
Scalar	94418.244	2160	Metric	2588.169 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	77015.305	2068				
Metric	89657.183	2114	Configural	12641.879 (46)	< .01	.003
Scalar	98108.681	2160	Metric	8451.497 (46)	< .01	.001

Appendix B.13b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	98752.423	2160	< .01	0.958	0.037
Model B-1	40679.803	2160	< .01	0.960	0.033
Model B-2	80280.423	2160	< .01	0.953	0.036
Model B-3	35866.572	2160	< .01	0.957	0.031
Model B-4	40737.236	2160	< .01	0.960	0.033
Model B-5	34986.399	2160	< .01	0.962	0.031
Model C	96412.066	2160	< .01	0.952	0.034
Model D	93599.712	2160	< .01	0.957	0.036
Model E	94418.244	2160	< .01	0.964	0.033
Model F	98108.681	2160	< .01	0.956	0.033

Appendix B.14a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	37314.410	2068				
Metric	38506.401	2114	Configural	1191.991 (46)	< .01	.000
Scalar	44483.010	2160	Metric	5976.609 (46)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	18571.581	2068				
Metric	20221.959	2114	Configural	1650.378 (46)	< .01	.001
Scalar	20875.847	2160	Metric	653.888 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	30441.823	2068				
Metric	35078.471	2114	Configural	4636.648 (46)	< .01	.001
Scalar	36819.342	2160	Metric	1740.870 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	17534.361	2068				
Metric	17791.556	2114	Configural	257.195 (46)	< .01	.000
Scalar	18399.719	2160	Metric	608.162 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	18001.839	2068				
Metric	19957.793	2114	Configural	1955.954 (46)	< .01	.001
Scalar	21012.517	2160	Metric	1054.724 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	17639.958	2068				
Metric	17735.461	2114	Configural	95.503 (46)	< .01	.000
Scalar	17822.776	2160	Metric	87.315 (46)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	35767.378	2068				
Metric	39249.591	2114	Configural	3482.213 (46)	< .01	.001
Scalar	44955.521	2160	Metric	5705.930 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	36685.569	2068				
Metric	39848.742	2114	Configural	3163.173 (46)	< .01	.001
Scalar	40105.854	2160	Metric	257.113 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	36944.840	2068				
Metric	39359.374	2114	Configural	2414.535 (46)	< .01	.001
Scalar	41609.312	2160	Metric	2249.938 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	34993.541	2068				
Metric	39720.647	2114	Configural	4727.106 (46)	< .01	.001
Scalar	45818.461	2160	Metric	6097.814 (46)	< .01	.002

Appendix B.14b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	44483.010	2160	< .01	0.980	0.022
Model B-1	20875.847	2160	< .01	0.981	0.021
Model B-2	36819.342	2160	< .01	0.977	0.022
Model B-3	18399.719	2160	< .01	0.982	0.020
Model B-4	21012.517	2160	< .01	0.982	0.020
Model B-5	17822.776	2160	< .01	0.983	0.019
Model C	44955.521	2160	< .01	0.980	0.020
Model D	40105.854	2160	< .01	0.981	0.021
Model E	41609.312	2160	< .01	0.985	0.019
Model F	45818.461	2160	< .01	0.981	0.020

Appendix B.15a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	51923.973	2068				
Metric	53374.054	2114	Configural	1450.081 (46)	< .01	.000
Scalar	57215.968	2160	Metric	3841.913 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	24553.604	2068				
Metric	25354.154	2114	Configural	800.550 (46)	< .01	.000
Scalar	26162.288	2160	Metric	808.134 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43941.130	2068				
Metric	45642.765	2114	Configural	1701.635 (46)	< .01	.000
Scalar	46780.427	2160	Metric	1137.662 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	23020.720	2068				
Metric	23325.360	2114	Configural	304.640 (46)	< .01	.000
Scalar	23646.081	2160	Metric	320.721 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	24766.121	2068				
Metric	25703.123	2114	Configural	937.002 (46)	< .01	.000
Scalar	26643.128	2160	Metric	940.004 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	23056.447	2068				
Metric	23120.850	2114	Configural	64.402 (46)	.038	.000
Scalar	23224.440	2160	Metric	103.590 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	48778.924	2068				
Metric	52531.621	2114	Configural	3752.697 (46)	< .01	.001
Scalar	57853.129	2160	Metric	5321.508 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	52081.658	2068				
Metric	53241.940	2114	Configural	1160.282 (46)	< .01	.000
Scalar	53724.454	2160	Metric	482.513 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	51790.117	2068				
Metric	52710.250	2114	Configural	920.133 (46)	< .01	.000
Scalar	54441.917	2160	Metric	1731.667 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	48522.307	2068				
Metric	51976.900	2114	Configural	3454.593 (46)	< .01	.001
Scalar	58394.447	2160	Metric	6417.548 (46)	< .01	.001

Appendix B.15b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	57215.968	2160	< .01	0.965	0.028
Model B-1	26162.288	2160	< .01	0.964	0.026
Model B-2	46780.427	2160	< .01	0.964	0.027
Model B-3	23646.081	2160	< .01	0.964	0.026
Model B-4	26643.128	2160	< .01	0.962	0.027
Model B-5	23224.440	2160	< .01	0.966	0.025
Model C	57853.129	2160	< .01	0.965	0.024
Model D	53724.454	2160	< .01	0.966	0.026
Model E	54441.917	2160	< .01	0.971	0.023
Model F	58394.447	2160	< .01	0.966	0.024

Appendix B.17a. Global Model Fit Indices of Measurement Invariance Tests for Algebra I

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	31880.312	2068				
Metric	32905.318	2114	Configural	1025.006 (46)	< .01	.000
Scalar	36406.798	2160	Metric	3501.480 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	16142.158	2068				
Metric	17121.893	2114	Configural	979.735 (46)	< .01	.001
Scalar	17604.115	2160	Metric	482.221 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	26017.305	2068				
Metric	29072.903	2114	Configural	3055.599 (46)	< .01	.001
Scalar	30186.879	2160	Metric	1113.975 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	15857.962	2068				
Metric	16107.596	2114	Configural	249.634 (46)	< .01	.000
Scalar	16536.207	2160	Metric	428.611 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	16758.206	2068				
Metric	18161.297	2114	Configural	1403.091 (46)	< .01	.001
Scalar	18954.656	2160	Metric	793.359 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	15465.159	2068				
Metric	15518.118	2114	Configural	52.959 (46)	.224	.001
Scalar	15611.302	2160	Metric	93.185 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	30806.225	2068				
Metric	32582.106	2114	Configural	1775.881 (46)	< .01	.000
Scalar	34778.963	2160	Metric	2196.858 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	31432.467	2068				
Metric	33610.533	2114	Configural	2178.066 (46)	< .01	.001
Scalar	34009.880	2160	Metric	399.346 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	31591.441	2068				
Metric	32369.885	2114	Configural	778.443 (46)	< .01	.000
Scalar	33674.958	2160	Metric	1305.073 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	31292.754	2068				
Metric	32158.284	2114	Configural	865.530 (46)	< .01	.000
Scalar	33945.363	2160	Metric	1787.079 (46)	< .01	.000

Appendix B.17b. Global Model Fit Indices of Scalar Invariance Model for Algebra I

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	36406.798	2160	< .01	0.929	0.024
Model B-1	17604.115	2160	< .01	0.981	0.021
Model B-2	30186.879	2160	< .01	0.974	0.023
Model B-3	16536.207	2160	< .01	0.981	0.021
Model B-4	18954.656	2160	< .01	0.979	0.022
Model B-5	15611.302	2160	< .01	0.983	0.020
Model C	34778.963	2160	< .01	0.927	0.024
Model D	34009.880	2160	< .01	0.930	0.023
Model E	33674.958	2160	< .01	0.936	0.023
Model F	33945.363	2160	< .01	0.982	0.020

Appendix B.16a. Global Model Fit Indices of Measurement Invariance Tests for Geometry

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	53224.910	2068				
Metric	54454.502	2114	Configural	1229.591 (46)	< .01	.000
Scalar	57940.176	2160	Metric	3485.674 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	27938.686	2068				
Metric	28828.529	2114	Configural	889.843 (46)	< .01	.000
Scalar	29338.108	2160	Metric	509.579 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43389.350	2068				
Metric	47107.528	2114	Configural	3718.178 (46)	< .01	.001
Scalar	48016.657	2160	Metric	909.129 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	27092.053	2068				
Metric	27384.954	2114	Configural	292.902 (46)	< .01	.000
Scalar	27629.871	2160	Metric	244.916 (46)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	27989.160	2068				
Metric	29872.199	2114	Configural	1883.040 (46)	< .01	.001
Scalar	30830.067	2160	Metric	957.867 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26391.366	2068				
Metric	26454.861	2114	Configural	63.495 (46)	.045	.001
Scalar	26525.935	2160	Metric	71.074 (46)	.010	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	52492.263	2068				
Metric	53783.151	2114	Configural	1290.888 (46)	< .01	.000
Scalar	55970.200	2160	Metric	2187.049 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	51894.256	2068				
Metric	54836.635	2114	Configural	2942.380 (46)	< .01	.001
Scalar	55244.890	2160	Metric	408.255 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	53475.540	2068				
Metric	54059.554	2114	Configural	584.014 (46)	< .01	.000
Scalar	54692.029	2160	Metric	632.476 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	52971.934	2068				
Metric	53842.768	2114	Configural	870.834 (46)	< .01	.000
Scalar	55431.388	2160	Metric	1588.620 (46)	< .01	.000

Appendix B.16b. Global Model Fit Indices of Scalar Invariance Model for Geometry

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	57940.176	2160	< .01	0.930	0.038
Model B-1	29338.108	2160	< .01	0.950	0.033
Model B-2	48016.657	2160	< .01	0.929	0.035
Model B-3	27629.871	2160	< .01	0.947	0.034
Model B-4	30830.067	2160	< .01	0.951	0.032
Model B-5	26525.935	2160	< .01	0.954	0.031
Model C	55970.200	2160	< .01	0.942	0.031
Model D	55244.890	2160	< .01	0.934	0.036
Model E	54692.029	2160	< .01	0.944	0.032
Model F	55431.388	2160	< .01	0.955	0.027

Appendix B.18a. Global Model Fit Indices of Measurement Invariance Tests for Algebra II

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	17718.329	2068				
Metric	18874.760	2114	Configural	1156.431 (46)	< .01	.000
Scalar	20911.565	2160	Metric	2036.806 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	10607.551	2068				
Metric	11002.537	2114	Configural	394.986 (46)	< .01	.000
Scalar	11497.152	2160	Metric	494.615 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	15096.010	2068				
Metric	16321.688	2114	Configural	1225.678 (46)	< .01	.001
Scalar	17382.791	2160	Metric	1061.103 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	10235.537	2068				
Metric	10522.133	2114	Configural	286.595 (46)	< .01	.001
Scalar	10828.678	2160	Metric	306.545 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	10485.655	2068				
Metric	11186.459	2114	Configural	700.803 (46)	< .01	.001
Scalar	12270.158	2160	Metric	1083.699 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	10048.537	2068				
Metric	10133.402	2114	Configural	84.865 (46)	< .01	.001
Scalar	10216.730	2160	Metric	83.328 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	17849.740	2068				
Metric	18266.012	2114	Configural	416.272 (46)	< .01	.000
Scalar	19649.191	2160	Metric	1383.179 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	17496.271	2068				
Metric	18725.869	2114	Configural	1229.598 (46)	< .01	.001
Scalar	19370.546	2160	Metric	644.677 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	17883.903	2068				
Metric	18249.789	2114	Configural	365.886 (46)	< .01	.000
Scalar	18641.474	2160	Metric	391.685 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	18128.269	2068				
Metric	18473.548	2114	Configural	345.279 (46)	< .01	.000
Scalar	19474.733	2160	Metric	1001.185 (46)	< .01	.000

Appendix B.18b. Global Model Fit Indices of Scalar Invariance Model for Algebra II

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	20911.565	2160	< .01	0.979	0.017
Model B-1	11497.152	2160	< .01	0.986	0.014
Model B-2	17382.791	2160	< .01	0.978	0.017
Model B-3	10828.678	2160	< .01	0.985	0.015
Model B-4	12270.158	2160	< .01	0.986	0.015
Model B-5	10216.730	2160	< .01	0.988	0.013
Model C	19649.191	2160	< .01	0.986	0.013
Model D	19370.546	2160	< .01	0.982	0.016
Model E	18641.474	2160	< .01	0.986	0.013
Model F	19474.733	2160	< .01	0.989	0.011

Appendix C: Regression Model Parameter Estimates of Differential Growth across Subgroups – ELA

Parameter	2015_G3E to 2016_G4E			2015_G4E to 2016_G5E		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	2521.59	0.14	<.0001	2539.23	0.14	<.0001
Female vs.Male (β_{01})	1.42	0.13	<.0001	3.63	0.14	<.0001
Special Education Status vs. Non-SPED (β_{02})	-6.70	0.28	<.0001	-8.89	0.29	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-11.12	0.41	<.0001	-9.94	0.49	<.0001
Low income vs. Non-Low Income(β_{04})	-2.26	0.15	<.0001	-1.69	0.16	<.0001
Asian vs.White (β_{05})	3.37	0.49	<.0001	3.49	0.49	<.0001
Hispanic vs.White (β_{06})	-3.41	0.16	<.0001	-3.55	0.16	<.0001
African American vs.White (β_{07})	-3.44	0.34	<.0001	-5.99	0.34	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-1.40	1.06	0.1846	-0.69	1.23	0.5744
American Indian vs.White (β_{09})	-6.71	0.37	<.0001	-7.67	0.37	<.0001
Multiple vs.White (β_{010})	-0.44	0.42	0.2939	-1.59	0.45	0.0004
Slope (β_{10})	0.85	0.00	<.0001	0.92	0.00	<.0001
Female vs.Male (β_{11})	0.00	0.00	0.8830	0.03	0.00	<.0001
Special Education Status vs. Non-SPED (β_{12})	0.04	0.01	<.0001	-0.03	0.01	<.0001
Limited English Proficiency vs. Non-LEP (β_{13})	-0.19	0.01	<.0001	-0.20	0.01	<.0001
Low income vs. Non-Low Income (β_{14})	-0.04	0.01	<.0001	-0.03	0.01	<.0001
Asian vs.White (β_{15})	-0.01	0.01	0.3709	-0.04	0.01	0.0036
Hispanic vs.White (β_{16})	0.02	0.01	<.0001	-0.02	0.01	0.0003
African American vs.White (β_{17})	0.06	0.01	<.0001	-0.04	0.01	0.0013
Hawaiian/Pacific Islander vs.White (β_{18})	0.02	0.04	0.5411	0.00	0.05	0.9591
American Indian vs.White (β_{19})	-0.05	0.01	<.0001	-0.07	0.01	<.0001
Multiple vs.White (β_{110})	0.00	0.01	0.8149	0.00	0.02	0.9071

Parameter	2015 G5E to 2016 G6E			2015 G6E to 2016 G7E		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	2544.30	0.14	<.0001	2555.45	0.12	<.0001
Female vs.Male (β_{01})	0.82	0.13	<.0001	0.62	0.12	<.0001
Special Education Status vs. Non-SPED (β_{02})	-7.46	0.32	<.0001	-8.01	0.31	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-13.23	0.56	<.0001	-9.31	0.62	<.0001
Low income vs. Non-Low Income(β_{04})	-2.00	0.15	<.0001	-1.90	0.14	<.0001
Asian vs.White (β_{05})	4.30	0.46	<.0001	2.51	0.42	<.0001
Hispanic vs.White (β_{06})	-4.83	0.16	<.0001	-2.70	0.14	<.0001
African American vs.White (β_{07})	-5.42	0.33	<.0001	-2.81	0.29	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	1.19	1.23	0.3313	-3.26	1.14	0.0042
American Indian vs.White (β_{09})	-7.48	0.38	<.0001	-5.70	0.37	<.0001
Multiple vs.White (β_{010})	-1.96	0.45	<.0001	-1.41	0.43	0.0010
Slope (β_{10})	0.98	0.00	<.0001	0.80	0.00	<.0001
Female vs.Male (β_{11})	0.02	0.00	<.0001	-0.01	0.00	0.0018
Special Education Status vs. Non-SPED (β_{12})	-0.12	0.01	<.0001	0.00	0.01	0.8637
Limited English Proficiency vs. Non-LEP (β_{13})	-0.31	0.01	<.0001	-0.11	0.02	<.0001
Low income vs. Non-Low Income (β_{14})	-0.06	0.01	<.0001	0.00	0.01	0.7904
Asian vs.White (β_{15})	0.01	0.01	0.3688	0.01	0.01	0.3462
Hispanic vs.White (β_{16})	-0.04	0.01	<.0001	0.01	0.00	0.0962

African American vs.White (β_{17})	-0.04	0.01	0.0013	0.01	0.01	0.2224
Hawaiian/Pacific Islander vs.White (β_{18})	0.06	0.05	0.1736	0.01	0.04	0.8099
American Indian vs.White (β_{19})	-0.11	0.01	<.0001	0.01	0.01	0.2810
Multiple vs.White (β_{110})	0.02	0.02	0.1585	0.01	0.01	0.3085

Parameter	2015 G7E to 2016 G8E			2015 G8E to 2016 G9E		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	2557.47	0.13	<.0001	2568.20	0.13	<.0001
Female vs.Male (β_{01})	1.53	0.13	<.0001	2.65	0.13	<.0001
Special Education Status vs. Non-SPED (β_{02})	-8.45	0.33	<.0001	-7.16	0.39	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-4.25	0.54	<.0001	-2.46	0.49	<.0001
Low income vs. Non-Low Income(β_{04})	-0.66	0.15	<.0001	-4.07	0.17	<.0001
Asian vs.White (β_{05})	4.53	0.45	<.0001	4.34	0.46	<.0001
Hispanic vs.White (β_{06})	-2.38	0.15	<.0001	-4.10	0.15	<.0001
African American vs.White (β_{07})	-3.00	0.30	<.0001	-4.07	0.33	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-1.73	1.23	0.1590	-1.43	1.11	0.1982
American Indian vs.White (β_{09})	-5.18	0.38	<.0001	-4.06	0.41	<.0001
Multiple vs.White (β_{010})	-0.83	0.45	0.0679	-0.34	0.54	0.5256
Slope (β_{10})	0.91	0.00	<.0001	0.84	0.00	<.0001
Female vs.Male (β_{11})	-0.01	0.00	0.0195	-0.01	0.00	0.0351
Special Education Status vs. Non-SPED (β_{12})	-0.09	0.01	<.0001	-0.05	0.01	<.0001
Limited English Proficiency vs. Non-LEP (β_{13})	-0.09	0.01	<.0001	0.01	0.01	0.5765
Low income vs. Non-Low Income (β_{14})	-0.01	0.01	0.0779	-0.05	0.01	<.0001
Asian vs.White (β_{15})	-0.01	0.01	0.6154	0.04	0.01	0.0027
Hispanic vs.White (β_{16})	0.00	0.01	0.6914	-0.02	0.01	0.0012
African American vs.White (β_{17})	0.01	0.01	0.5551	0.01	0.01	0.4781
Hawaiian/Pacific Islander vs.White (β_{18})	-0.01	0.04	0.7772	0.04	0.04	0.3859
American Indian vs.White (β_{19})	-0.06	0.01	<.0001	-0.05	0.01	0.0010
Multiple vs.White (β_{110})	0.00	0.02	0.7872	0.00	0.02	0.9819

Parameter	2015 G9E to 2016 G10E			2015 G10E to 2016 G11E		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	2569.76	0.13	<.0001	2571.39	0.15	<.0001
Female vs.Male (β_{01})	-0.43	0.14	0.0021	1.57	0.16	<.0001
Special Education Status vs. Non-SPED (β_{02})	-5.72	0.40	<.0001	-8.01	0.48	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-4.23	0.48	<.0001	-2.70	0.62	<.0001
Low income vs. Non-Low Income(β_{04})	-3.51	0.18	<.0001	-4.68	0.22	<.0001
Asian vs.White (β_{05})	2.66	0.47	<.0001	2.58	0.52	<.0001
Hispanic vs.White (β_{06})	-3.66	0.16	<.0001	-3.86	0.18	<.0001
African American vs.White (β_{07})	-4.45	0.33	<.0001	-4.48	0.40	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-3.40	1.38	0.0139	-2.88	1.45	0.0476
American Indian vs.White (β_{09})	-5.15	0.41	<.0001	-4.25	0.47	<.0001
Multiple vs.White (β_{010})	-0.30	0.56	0.5959	-0.43	0.64	0.5073
Intercept (β_{00})	0.86	0.01	<.0001	0.91	0.01	<.0001
Female vs.Male (β_{01})	-0.01	0.01	0.0638	0.00	0.01	0.9377
Special Education Status vs. Non-SPED (β_{02})	-0.06	0.01	<.0001	-0.09	0.01	<.0001

Limited English Proficiency vs. Non-LEP (β_{03})	-0.02	0.02	0.2390	-0.05	0.02	0.0261
Low income vs. Non-Low Income(β_{04})	-0.07	0.01	<.0001	-0.03	0.01	0.0021
Asian vs.White (β_{05})	0.06	0.01	<.0001	0.08	0.02	<.0001
Hispanic vs.White (β_{06})	-0.03	0.01	<.0001	-0.03	0.01	0.0001
African American vs.White (β_{07})	-0.03	0.01	0.0323	-0.01	0.02	0.3888
Hawaiian/Pacific Islander vs.White (β_{08})	0.07	0.05	0.1502	0.01	0.06	0.8333
American Indian vs.White (β_{09})	-0.10	0.02	<.0001	-0.11	0.02	<.0001
Multiple vs.White (β_{010})	0.00	0.02	0.8904	0.00	0.02	0.9491

Note: G3E through G11E refers grade 3 ELA to grade 11 ELA. The significance of the effect is less than p value of .05.

Appendix C cont. Regression Model Parameter Estimates of Differential Growth across Subgroups – Math

Parameter	2015 G3M to 2016 G4M			2015 G4M to 2016 G5M		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	3558.95	0.17	<.0001	3590.54	0.18	<.0001
Female vs.Male (β_{01})	-0.01	0.16	0.9286	3.55	0.17	<.0001
Special Education Status vs. Non-SPED (β_{02})	-8.44	0.33	<.0001	-10.00	0.35	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-9.16	0.43	<.0001	-6.60	0.56	<.0001
Low income vs. Non-Low Income(β_{04})	-0.36	0.18	0.0498	-1.40	0.19	<.0001
Asian vs.White (β_{05})	3.79	0.65	<.0001	7.97	0.64	<.0001
Hispanic vs.White (β_{06})	-6.32	0.19	<.0001	-2.11	0.20	<.0001
African American vs.White (β_{07})	-9.13	0.42	<.0001	-4.04	0.43	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-3.93	1.30	0.0025	-1.11	1.51	0.4595
American Indian vs.White (β_{09})	-9.11	0.44	<.0001	-6.60	0.46	<.0001
Multiple vs.White (β_{010})	-3.15	0.50	<.0001	-0.67	0.55	0.2224
Slope (β_{10})	0.76	0.00	<.0001	0.82	0.00	<.0001
Female vs.Male (β_{11})	-0.01	0.00	0.0099	-0.02	0.00	<.0001
Special Education Status vs. Non-SPED (β_{12})	0.10	0.00	<.0001	0.08	0.01	<.0001
Limited English Proficiency vs. Non-LEP (β_{13})	-0.05	0.00	<.0001	-0.01	0.01	0.5093
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.01	0.01	0.1949
Asian vs.White (β_{15})	-0.01	0.00	0.2898	0.01	0.01	0.4984
Hispanic vs.White (β_{16})	0.00	0.00	0.6600	0.02	0.01	<.0001
African American vs.White (β_{17})	0.03	0.00	0.0012	0.05	0.01	0.0001
Hawaiian/Pacific Islander vs.White (β_{18})	-0.01	0.00	0.7891	-0.04	0.04	0.4039
American Indian vs.White (β_{19})	0.00	0.00	0.9644	0.00	0.01	0.9941
Multiple vs.White (β_{110})	0.01	0.00	0.3142	0.01	0.01	0.3647

Parameter	2015 G5M to 2016 G6M			2015 G6M to 2016 G7M		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	3620.97	0.17	<.0001	3637.50	0.14	<.0001
Female vs.Male (β_{01})	2.07	0.16	<.0001	0.28	0.14	0.0448
Special Education Status vs. Non-SPED (β_{02})	-11.02	0.37	<.0001	-10.27	0.34	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-12.58	0.59	<.0001	-11.15	0.59	<.0001
Low income vs. Non-Low Income(β_{04})	-1.22	0.19	<.0001	-2.74	0.16	<.0001
Asian vs.White (β_{05})	2.25	0.61	0.0002	2.51	0.55	<.0001
Hispanic vs.White (β_{06})	-5.52	0.19	<.0001	-3.82	0.16	<.0001
African American vs.White (β_{07})	-6.41	0.42	<.0001	-5.22	0.35	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-0.86	1.51	0.5666	-2.60	1.30	0.0455

American Indian vs.White (β_{09})	-8.41	0.45	<.0001	-8.51	0.38	<.0001
Multiple vs.White (β_{010})	-2.11	0.55	0.0001	-0.89	0.49	0.0676
Slope (β_{10})	0.83	0.00	<.0001	0.81	0.00	<.0001
Female vs.Male (β_{11})	0.00	0.00	0.4006	-0.01	0.00	0.0122
Special Education Status vs. Non-SPED (β_{12})	0.11	0.01	<.0001	-0.01	0.01	0.5219
Limited English Proficiency vs. Non-LEP (β_{13})	-0.07	0.01	<.0001	-0.14	0.02	<.0001
Low income vs. Non-Low Income (β_{14})	0.01	0.01	0.2319	0.01	0.01	0.0479
Asian vs.White (β_{15})	0.02	0.01	0.1103	0.01	0.01	0.6488
Hispanic vs.White (β_{16})	0.01	0.01	0.0058	-0.01	0.00	0.0385
African American vs.White (β_{17})	0.07	0.01	<.0001	0.02	0.01	0.0349
Hawaiian/Pacific Islander vs.White (β_{18})	0.09	0.04	0.0219	0.01	0.04	0.7645
American Indian vs.White (β_{19})	0.02	0.01	0.0702	-0.04	0.01	0.0002
Multiple vs.White (β_{110})	0.03	0.01	0.0515	0.01	0.01	0.5148

Parameter	2015 G7M to 2016 G8M			2015 G8M to 2016 AlgI		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	3653.40	0.17	<.0001	3666.15	0.18	<.0001
Female vs.Male (β_{01})	2.48	0.16	<.0001	2.44	0.17	<.0001
Special Education Status vs. Non-SPED (β_{02})	-7.96	0.36	<.0001	-7.29	0.39	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-7.65	0.61	<.0001	-3.90	0.55	<.0001
Low income vs. Non-Low Income(β_{04})	-1.55	0.18	<.0001	-3.87	0.21	<.0001
Asian vs.White (β_{05})	5.31	0.67	<.0001	3.79	0.71	<.0001
Hispanic vs.White (β_{06})	-2.07	0.19	<.0001	-2.49	0.19	<.0001
African American vs.White (β_{07})	-2.06	0.38	<.0001	-2.36	0.40	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	0.75	1.53	0.6267	-1.74	1.47	0.2389
American Indian vs.White (β_{09})	-4.85	0.42	<.0001	-2.65	0.46	<.0001
Multiple vs.White (β_{010})	-1.00	0.58	0.0843	-1.54	0.73	0.0333
Slope (β_{10})	0.86	0.00	<.0001	0.76	0.01	<.0001
Female vs.Male (β_{11})	0.00	0.00	0.6631	-0.01	0.01	0.0114
Special Education Status vs. Non-SPED (β_{12})	-0.04	0.01	<.0001	-0.13	0.01	<.0001
Limited English Proficiency vs. Non-LEP (β_{13})	-0.12	0.02	<.0001	-0.07	0.02	<.0001
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.03	0.01	<.0001
Asian vs.White (β_{15})	0.03	0.01	0.0546	0.04	0.02	0.0666
Hispanic vs.White (β_{16})	0.01	0.01	0.0981	-0.05	0.01	<.0001
African American vs.White (β_{17})	-0.01	0.01	0.6058	-0.06	0.01	<.0001
Hawaiian/Pacific Islander vs.White (β_{18})	0.00	0.04	0.9952	-0.07	0.06	0.1891
American Indian vs.White (β_{19})	-0.04	0.01	0.0065	-0.08	0.02	<.0001
Multiple vs.White (β_{110})	0.03	0.02	0.1548	-0.03	0.03	0.2957

Parameter	2015 AlgI to 2016 Geo			2015 Geo to 2016 AlgII		
	Estimate	Standard Error	p value	Estimate	Standard Error	p value
Intercept (β_{00})	3688.61	0.18	<.0001	3699.15	0.19	<.0001
Female vs.Male (β_{01})	0.25	0.18	0.1716	1.37	0.20	<.0001
Special Education Status vs. Non-SPED (β_{02})	-8.23	0.50	<.0001	-7.22	0.61	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-3.27	0.56	<.0001	1.28	0.70	0.0676
Low income vs. Non-Low Income(β_{04})	-3.49	0.25	<.0001	-4.97	0.28	<.0001

Asian vs.White (β_{05})	0.90	0.63	0.1512	4.86	0.64	<.0001
Hispanic vs.White (β_{06})	-5.06	0.21	<.0001	-4.00	0.23	<.0001
African American vs.White (β_{07})	-5.83	0.47	<.0001	-3.35	0.52	<.0001
Hawaiian/Pacific Islander vs.White (β_{08})	-5.64	1.76	0.0014	-4.30	1.90	0.0233
American Indian vs.White (β_{09})	-4.27	0.53	<.0001	-7.36	0.56	<.0001
Multiple vs.White (β_{010})	-2.65	0.74	0.0003	-0.30	0.79	0.7065
Slope (β_{10})	0.84	0.00	<.0001	0.79	0.01	<.0001
Female vs.Male (β_{11})	-0.01	0.01	0.0887	-0.01	0.01	0.3208
Special Education Status vs. Non-SPED (β_{12})	-0.08	0.01	<.0001	-0.10	0.02	<.0001
Limited English Proficiency vs. Non-LEP (β_{13})	0.02	0.02	0.2392	0.08	0.02	0.0004
Low income vs. Non-Low Income (β_{14})	-0.05	0.01	<.0001	-0.07	0.01	<.0001
Asian vs.White (β_{15})	0.03	0.01	0.0369	0.09	0.02	<.0001
Hispanic vs.White (β_{16})	-0.05	0.01	<.0001	-0.10	0.01	<.0001
African American vs.White (β_{17})	-0.04	0.02	0.0106	-0.08	0.02	<.0001
Hawaiian/Pacific Islander vs.White (β_{18})	0.04	0.06	0.4755	0.01	0.06	0.9300
American Indian vs.White (β_{19})	-0.07	0.02	<.0001	-0.15	0.02	<.0001
Multiple vs.White (β_{110})	0.00	0.02	0.9952	0.00	0.02	0.8723

Note: G3M through G8M refers grade 3 Math to grade 8 Math. AlgI and AlgII refers Algebra I and II, respectively.
Geo refers Geometry. The significance of the effect is less than p value of .05.

Appendix D.1 – Student Participation by Demographic Subgroup – Fall 2015 Online Administration_online

Group	G9E	G10E	G11E	AlgI	Geo	AlgII
All students	2934	3304	5097	6084	4909	4756
Female	1354	1502	2484	2923	2258	2398
Male	1580	1802	2613	3161	2651	2358
African American	187	189	351	438	405	292
Asian	59	79	119	192	90	162
Native Hawaiian/Pacific	11	14	17	29	14	24
Hispanic/Latino	1285	1348	2340	2500	2356	1826
American Indian or Alaskan	268	228	289	234	194	233
White	1059	1331	1867	2563	1705	2078
Multiple	65	115	114	128	145	141
Limited English Proficiency	72	39	62	97	75	62
Special Education	273	275	383	270	379	206

Appendix D.2 – Student Participation by Demographic Subgroup – Fall 2015 Online Administration_paper

Group	G9E	G10E	G11E	AlgI	Geo	AlgII
All students	864	739	996	1275	978	656
Female	411	351	481	591	472	340
Male	453	388	515	684	506	316
African American	50	46	53	94	45	46
Asian	9	6	7	20	20	13
Native Hawaiian/Pacific	3	3	4	6	5	2
Hispanic/Latino	426	365	546	566	455	334
American Indian or Alaskan	38	27	41	51	28	12
White	317	267	331	506	393	237
Multiple	18	18	11	27	31	12
Limited English Proficiency	13	10	17	21	16	13
Special Education	77	91	140	179	138	53

Appendix E. Equations and Formula for Estimating Reliability

E.1 Standard Error Formula

For the AzMERIT assessments scored using MLE, according to Masters (1982), the asymptotic estimate of the standard error for ability θ is given by

$$SE(\theta) = \left[\sum_{i=1}^N \sum_{x_i=0}^{m_i} x_i^2 P(X_i = x_i | \theta) - \sum_{i=1}^N \left[\sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right]^2 \right]^{-\frac{1}{2}},$$

which is further placed onto the reporting scale by the following transformation:

$$SE_{vs} = a \times SE(\theta),$$

where a is the slope of the scaling constants that take θ to the reporting scale. For both ELA and Mathematics tests, $a = 30$.

E.2 Student Classification Consistency Formula

For a student with estimated ability $\hat{\theta}$ and associated standard error $se(\hat{\theta})$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $se(\hat{\theta})$, that is, $\hat{\theta} \sim N(\theta, se(\hat{\theta})^2)$. The probability of the true score *at or above* the cut score θ_c is estimated as

$$P(\theta \geq \theta_c) = P\left(\frac{\theta - \hat{\theta}}{se(\hat{\theta})} \geq \frac{\theta_c - \hat{\theta}}{se(\hat{\theta})}\right) = P\left(\frac{\hat{\theta} - \theta}{se(\hat{\theta})} < \frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right),$$

where $\Phi(\cdot)$ is the cumulative function of standard normal distribution. Similarly, the probability of the true score being *below* the cut score is estimated as

$$P(\theta < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right).$$

E.2.1 Classification Accuracy Formula

The probability of a student with true ability θ being classified *at or above* the cut score θ_c , given the student's item scores $\mathbf{x} = (x_1, \dots, x_N)$, can be estimated as

$$P(\theta \geq \theta_c | \mathbf{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta},$$

where the likelihood function is

$$L(\theta | \mathbf{x}) = \prod_{i=1}^N P(x_i | \theta),$$

and $P(x_i|\theta)$ is calculated from the Rasch model or partial credit model based on the estimated item parameters.

Similarly, we can estimate the probability of *below* the cut score as:

$$P(\theta < \theta_c | \mathbf{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta}$$

Mathematically, we have

$$\begin{aligned} N_{11} &= \sum_{i \in N_1} P(\theta_i \geq \theta_c | \mathbf{x}), \\ N_{01} &= \sum_{i \in N_1} P(\theta_i < \theta_c | \mathbf{x}), \\ N_{10} &= \sum_{i \in N_0} P(\theta_i \geq \theta_c | \mathbf{x}), \text{ and} \\ N_{00} &= \sum_{i \in N_0} P(\theta_i < \theta_c | \mathbf{x}), \end{aligned}$$

where N_1 consists of the students with estimated $\hat{\theta}_i$ being *at and above* the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being *below* the cut score. The accuracy index is then computed as:

$$\frac{N_{11} + N_{00}}{N_1 + N_0}.$$

E.2.2 Classification Consistency Formula

To estimate the consistency, we assume the students are tested twice independently; hence, the probability of the student being classified as *at or above* the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

Similarly, the probability of consistency for *at or above* the cut score is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c | \mathbf{x}) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

The probability of consistency for *below* the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c | \mathbf{x}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

The probability of inconsistency is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 < \theta_c | \mathbf{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta \int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta \right]^2}, \text{ and}$$

$$P(\theta_1 < \theta_c, \theta_2 \geq \theta_c | \mathbf{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta \int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta \right]^2}.$$

The consistent index is computed as $\frac{N_{11} + N_{00}}{N}$, where

$$N_{11} = \sum_{i \in N} P(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c | \mathbf{x}),$$

$$N_{01} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} \geq \theta_c | \mathbf{x}),$$

$$N_{10} = \sum_{i \in N} P(\theta_i \geq \theta_c, \theta_{i,2} < \theta_c | \mathbf{x}),$$

$$N_{00} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} < \theta_c | \mathbf{x}), \text{ and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

Appendix F: Standard Errors of Measurement – Fall 2015 Administration

Test	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	Overall
Grade 9 ELA	9.62	8.81	9.52	11.32	9.45
Grade 10 ELA	9.11	8.20	8.71	10.83	8.93
Grade 11 ELA	9.46	8.73	9.10	10.54	9.32
Algebra I	10.96	9.48	9.75	13.60	10.73
Geometry	12.80	10.53	10.16	11.31	11.77
Algebra II	13.79	10.96	10.43	11.23	12.24

Appendix G.1 – Number of Participating Students by Demographic Subgroups – ELA Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	62244	61069	60212	57639	58050	57350	51007	46818	41488
African American	3192	3194	3157	3120	3083	3077	2558	2373	1919
Asian	1458	1386	1384	1347	1329	1369	1231	1200	1109
Native Hawaiian/Pacific	235	244	208	170	155	144	205	149	143
Hispanic/Latino	29158	28182	27559	25946	25415	25172	21357	19155	16529
American Indian or Alaskan	3385	3412	3466	3166	3105	3048	3046	2818	2556
White	22945	22938	22897	22530	23711	23374	21841	20447	18653
Multiple	1871	1713	1541	1360	1252	1166	769	676	579
Female	30424	29986	29628	28073	28414	28138	24953	23085	20502
Male	31820	31083	30584	29566	29636	29212	26054	23733	20986
Limited English Proficiency	6505	6371	4921	3598	2998	2504	2394	2096	1465
Special Education	6797	7218	7236	6580	6085	5734	4395	3809	3127
Free/Reduced Lunch	26525	25856	25061	23567	22826	21773	18404	16146	13341
Accommodation	11724	11273	9492	7824	6533	5565	2495	1854	1422

Appendix G.2 – Number of Participating Students by Demographic Subgroups – ELA Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	25549	25256	25213	27012	26088	25429	29123	26585	23346
African American	1245	1238	1273	1398	1404	1478	1747	1622	1412
Asian	964	936	1082	1195	1090	1040	1192	1102	1011
Native Hawaiian/Pacific	78	105	69	102	90	84	83	58	74
Hispanic/Latino	11169	11023	10613	11623	11708	11450	14024	12334	10673
American Indian or Alaskan	969	921	960	1127	896	1015	854	756	597
White	10278	10212	10521	10904	10235	9771	10644	10136	9091
Multiple	832	809	687	658	649	579	528	522	452
Female	12710	12466	12474	13117	13004	12619	14535	13422	11706
Male	12825	12778	12731	13890	13068	12799	14540	13118	11610
Unknown Gender	14	12	8	5	16	11	48	45	30
Limited English Proficiency	1986	1940	1549	1318	1107	843	1093	520	290
Special Education	2784	2965	2968	2969	2818	2819	2551	2215	1869
Free/Reduced Lunch	11222	10999	10999	11899	11973	11659	11599	10341	8850
Accommodation	1715	1744	1439	1403	1079	905	616	337	227

Appendix G.3 – Number of Participating Students by Demographic Subgroups – Mathematics Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geomet ry	Algebra II
All students	62594	61395	60452	57874	57471	49787	53725	45784	39250
African American	3229	3241	3177	3141	3096	2898	2727	2161	1862
Asian	1465	1392	1386	1342	1246	947	1175	1203	1048
Native Hawaiian/Pacific	236	244	210	170	155	130	180	165	126
Hispanic/Latino	29321	28336	27651	26072	25421	23423	22997	18622	15576
American Indian or Alaskan	3416	3427	3483	3191	3178	2876	3567	2868	2346
White	23041	23034	22993	22589	23151	18504	22249	20079	17754
Multiple	1886	1721	1552	1369	1224	1009	830	686	538
Female	30576	30125	29723	28168	28201	24257	26023	22664	19805
Male	32018	31270	30729	29706	29270	25530	27702	23120	19445
Limited English Proficiency	6615	6450	4972	3639	3033	2310	2595	2135	1336
Special Education	6906	7287	7310	6621	6165	5729	4834	3286	2255
Free/Reduced Lunch	26077	25874	24986	23452	22820	20281	19568	16090	12676
Accommodation	11924	11518	9770	8083	6692	5652	2725	1665	1099

Appendix G.4 – Number of Participating Students by Demographic Subgroups – Mathematics Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geomet ry	Algebra II
All students	25709	25316	25267	26801	24358	20071	28898	25870	21650
African American	1259	1250	1277	1402	1341	1306	1779	1501	1334
Asian	970	949	1074	1011	638	468	1227	1153	903
Native Hawaiian/Pacific	80	104	69	101	88	70	81	57	66
Hispanic/Latino	11229	11019	10647	11666	11519	10167	14031	12020	10036
American Indian or Alaskan	988	926	968	1142	896	927	886	727	574
White	10330	10234	10535	10809	9244	6655	10320	9883	8255
Multiple	840	822	689	664	616	458	503	473	439
Female	12773	12482	12498	13005	12106	9846	14402	12988	10963
Male	12923	12822	12761	13790	12236	10206	14428	12829	10645
Unknown Gender	13	12	8	6	16	19	68	53	42
Limited English Proficiency	2007	1952	1560	1328	1110	843	1152	722	326
Special Education	2825	2978	2989	3007	2828	2727	2582	2033	1321
Free/Reduced Lunch	11080	10916	10946	11863	11867	10249	12292	10018	8632
Accommodation	1594	1616	1236	1226	919	750	343	221	148

Appendix H Operational Item Parameter
Estimates - Spring 2016 Administration

Appendix H.1—Spring 16 Operational Item Parameter Estimates — Grade 3 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13021_e	ER	-1.13009	2.10828	4.15506	1.71108
2	13021_o	ER	-1.10203	1.71104	4.1758	1.59494
3	13021_c	ER	-1.31327	-0.98811		-1.15069
4	13025_e	ER	-1.28491	2.22063	3.90277	1.61283
5	13025_o	ER	-0.6838	1.77292	4.29041	1.79318
6	13025_c	ER	-1.58499	-0.47921		-1.0321
7	13026_e	ER	-1.11168	1.96713	4.20208	1.68584
8	13026_o	ER	-1.02194	1.83366	4.15705	1.65626
9	13026_c	ER	-1.53632	-0.61273		-1.07453
10	13022_e	ER	-1.50579	1.76305	4.40933	1.55553
11	13022_o	ER	-1.67709	1.56754	4.24577	1.37874
12	13022_c	ER	-1.75863	-0.83246		-1.29555
13	13024_e	ER	-1.10833	2.26727	4.05905	1.73933
14	13024_o	ER	-1.24303	2.04797	4.13136	1.64543
15	13024_c	ER	-1.6353	-0.94878		-1.29204
16	13023_e	ER	-1.15476	2.25355	4.12678	1.74186
17	13023_o	ER	-1.34689	2.14582	3.94719	1.58204
18	13023_c	ER	-1.69753	-0.82307		-1.2603
19	9698	MC4	-1.11745			-1.11745
20	9691	MC4	-0.37757			-0.37757
21	9687	HT	1.38239			1.38239
22	9700	MC4	1.23623			1.23623
23	9690	MC4	-1.209			-1.209
24	9697	HT	-0.26674			-0.26674
25	9694	EBSR4	1.70532			1.70532
26	9692	MC4	-0.60073			-0.60073
27	9699	MC4	-0.30132			-0.30132
28	8708	ETC	-1.85419			-1.85419
29	8710	ETC	-1.27986			-1.27986
30	8711	ETC	-0.1057			-0.1057
31	9359	MC4	0.28114			0.28114
32	9356	HT	1.30318			1.30318
33	9357	MC4	0.82249			0.82249
34	9353	MC4	0.10251			0.10251
35	9355	MC4	-0.3629			-0.3629
36	10268	MC4	0.2005			0.2005
37	9330	MC4	-0.2646			-0.2646
38	9336	MC4	-0.0956			-0.0956

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9331	MC4	0.62573			0.62573
40	9335	MC4	-0.64679			-0.64679
41	9333	MC4	0.30718			0.30718
42	835	MC4	-1.54772			-1.54772
43	837	MC4	-0.18142			-0.18142
44	9402	MS5	1.66101			1.66101
45	9398	MC4	-0.0567			-0.0567
46	9401	EBSR4	1.0949			1.0949
47	9347	MC4	0.27951			0.27951
48	9349	MC4	0.81415			0.81415
49	9404	MC4	-0.26242			-0.26242
50	9400	MC4	0.49505			0.49505
51	9414	MC4	-0.03304			-0.03304
52	9422	MC4	-0.87555			-0.87555
53	9418	MC4	0.70984			0.70984
54	10632	MC4	0.48179			0.48179
55	10634	MC4	1.17725			1.17725
56	9377	ETC	0.18504			0.18504
57	9378	ETC	-1.0865	-0.3551		-0.7208
58	9379	ETC	-1.03119			-1.03119
59	9380	ETC	-0.87865			-0.87865

Appendix H.2—Spring 16 Operational Item Parameter Estimates — Grade 4 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	9395	MC4	-0.2812			-0.2812
2	9392	MC4	-0.68057			-0.68057
3	9426	HT	0.92555			0.92555
4	10266	MC4	-0.74849			-0.74849
5	9391	MC4	0.26422			0.26422
6	1122	MC4	1.3626			1.3626
7	1124	MS4	-0.66063			-0.66063
8	9598	MC4	-0.0095			-0.0095
9	9595	MC4	-0.42531			-0.42531
10	9597	MC4	-0.09234			-0.09234
11	9596	MC4	0.65108			0.65108
12	9603	MS6	2.72225			2.72225
13	10644	ETC	-0.05707			-0.05707
14	10645	ETC	-2.09722	0.11686		-0.99018
15	10647	ETC	-0.85463			-0.85463
16	10263	MC4	-0.72314			-0.72314
17	9425	MC4	-0.64738			-0.64738
18	9382	MC4	-1.1843			-1.1843
19	9387	MC4	1.48043			1.48043
20	9386	MS6	0.46331			0.46331
21	9388	HT	-0.33297			-0.33297
22	9397	EBSR4	0.81106			0.81106
23	9389	MC4	0.04148			0.04148
24	9616	MC4	-0.19377			-0.19377
25	9390	MC4	-0.88945			-0.88945
26	9903	HT	0.55857			0.55857
27	10278	EBSR4	2.12205			2.12205
28	10277	MC4	1.11434			1.11434
29	9899	MC4	0.52233			0.52233
30	9902	MC4	-1.2141			-1.2141
31	9906	MC4	-0.53878			-0.53878
32	9900	HT	2.72673			2.72673
33	9446	MC4	-0.77651			-0.77651
34	9439	MS6	2.08491			2.08491
35	9437	MC4	-0.03208			-0.03208
36	9451	HT	1.17836			1.17836
37	9450	MC4	-0.16915			-0.16915
38	9438	MS6	1.75326			1.75326

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9428	ETC	0.79794			0.79794
40	9429	ETC	-2.21688	0.57637		-0.82026
41	9431	ETC	0.47241			0.47241
42	13094_e	ER	1.87077	4.87517	6.53587	4.42727
43	13094_o	ER	1.25092	4.69117	6.1041	4.0154
44	13094_c	ER	-0.79626	1.71384		0.45879
45	13095_e	ER	1.87872	4.8299	4.92767	3.87876
46	13095_o	ER	0.47386	4.35392	4.92003	3.24927
47	13095_c	ER	-1.45942	1.23827		-0.11058
48	13119_e	ER	1.54979	4.4619	6.46226	4.15798
49	13119_o	ER	0.64594	4.34742	5.35763	3.45033
50	13119_c	ER	-1.65829	1.05358		-0.30236
51	13121_e	ER	1.87054	4.48532	7.08457	4.48014
52	13121_o	ER	0.80218	4.25587	8.11115	4.38973
53	13121_c	ER	-1.70071	1.64402		-0.02835
54	13118_e	ER	1.69157	4.56961	4.98225	3.74781
55	13118_o	ER	0.73557	4.29715	4.84718	3.2933
56	13118_c	ER	-1.60239	1.2628		-0.1698
57	13120_e	ER	1.98955	5.05842	4.4995	3.84916
58	13120_o	ER	0.90148	4.27097	6.19001	3.78749
59	13120_c	ER	-1.25847	1.5145		0.12802

Appendix H.3—Spring 16 Operational Item Parameter Estimates — Grade 5 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13246_e	ER	0.33484	4.37175	6.34064	3.68241
2	13246_o	ER	-0.55282	3.33662	5.17957	2.65446
3	13246_c	ER	-2.03231	0.35683		-0.83774
4	13236_e	ER	-0.76876	3.70295	4.83054	2.58824
5	13236_o	ER	-0.83522	2.91473	5.00997	2.36316
6	13236_c	ER	-2.13985	-0.09297		-1.11641
7	13237_e	ER	0.76049	4.48983	5.72602	3.65878
8	13237_o	ER	-0.3275	3.62724	5.71883	3.00619
9	13237_c	ER	-1.61986	0.18987		-0.715
10	13238_e	ER	-0.90788	3.36594	4.06761	2.17522
11	13238_o	ER	-1.30216	2.84958	4.12669	1.89137
12	13238_c	ER	-2.13488	0.01614		-1.05937
13	13239_e	ER	0.02633	3.71815	4.71382	2.81943
14	13239_o	ER	-1.10817	3.18175	4.64235	2.23864
15	13239_c	ER	-2.15823	0.01608		-1.07108
16	13247_e	ER	-0.45605	3.2947	4.52191	2.45352
17	13247_o	ER	-1.00143	2.7942	4.59779	2.13019
18	13247_c	ER	-2.16137	0.43573		-0.86282
19	9305	MC4	-1.06814			-1.06814
20	9303	MC4	0.14403			0.14403
21	9304	MS6	0.5267			0.5267
22	9302	MC4	0.397			0.397
23	10264	HT	2.62702			2.62702
24	9290	ETC	-0.70663			-0.70663
25	9291	ETC	-1.14142	0.69588		-0.22277
26	9292	ETC	-1.08057	0.56956		-0.25551
27	9754	MS5	3.23528			3.23528
28	9755	MS6	1.2478			1.2478
29	9751	MC4	-0.23706			-0.23706
30	9757	MC4	-0.08355			-0.08355
31	9752	HT	1.19823			1.19823
32	9767	EBSR4	0.0675			0.0675
33	9753	MS6	0.81962			0.81962
34	9312	HT	2.72402			2.72402
35	9308	MC4	-0.32336			-0.32336
36	9307	MC4	-0.67813			-0.67813
37	9299	MC4	-0.1376			-0.1376
38	9298	MC4	-0.82066			-0.82066

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9311	HT	2.10824			2.10824
40	9306	MC4	-0.33688			-0.33688
41	9320	EBSR4	1.17902			1.17902
42	9783	MC4	-0.86303			-0.86303
43	9784	MC4	0.73684			0.73684
44	9842	HT	0.63012			0.63012
45	9833	MC4	0.8069			0.8069
46	9808	MC4	0.19637			0.19637
47	9809	MC4	0.24291			0.24291
48	9758	HT	-0.48322			-0.48322
49	9762	MC4	0.05262			0.05262
50	9765	MC4	1.41073			1.41073
51	9763	MC4	0.47559			0.47559
52	9600	HT	0.29219			0.29219
53	9599	HT	0.11816			0.11816
54	9294	MC4	-1.04402			-1.04402
55	9309	HT	3.39654			3.39654
56	9601	MS5	-0.19919			-0.19919
57	10659	ETC	-0.82784			-0.82784
58	10661	ETC	-2.20002	-0.66798		-1.434
59	10662	ETC	-1.01469			-1.01469

Appendix H.4—Spring 16 Operational Item Parameter Estimates — Grade 6 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13307_e	ER	0.4746	3.03111	4.52238	2.67603
2	13307_o	ER	-0.54732	2.76978	4.2732	2.16522
3	13307_c	ER	-2.10835	-0.10872		-1.10854
4	13308_e	ER	1.2363	3.33177	5.17142	3.2465
5	13308_o	ER	0.1599	3.083	4.37256	2.53849
6	13308_c	ER	-2.10091	-0.21027		-1.15559
7	13309_e	ER	0.33264	2.78999	4.58425	2.56896
8	13309_o	ER	-0.4541	2.19863	4.51841	2.08765
9	13309_c	ER	-2.8389	0.01884		-1.41003
10	13304_e	ER	-0.40648	2.83619	5.19328	2.541
11	13304_o	ER	-0.7602	2.34515	4.23866	1.9412
12	13304_c	ER	-2.38004	-0.22638		-1.30321
13	13305_e	ER	0.79496	3.84657	4.56174	3.06776
14	13305_o	ER	0.0947	3.01629	4.70018	2.60372
15	13305_c	ER	-1.96936	-0.65918		-1.31427
16	13306_e	ER	0.07056	3.38151	6.19445	3.21551
17	13306_o	ER	-0.59192	2.51057	5.24377	2.38747
18	13306_c	ER	-2.06099	-0.58055		-1.32077
19	9083	ETC	-2.37343	0.69063		-0.8414
20	9084	ETC	-0.529	1.94599		0.7085
21	9145	MC4	-0.95899			-0.95899
22	9143	MC4	0.38775			0.38775
23	9139	MC4	-0.90857			-0.90857
24	9140	MC4	0.29413			0.29413
25	10265	MC4	-1.04948			-1.04948
26	9142	HT	1.591			1.591
27	9807	EBSR4	1.168			1.168
28	9797	HT	0.97683			0.97683
29	9799	MC4	-0.16807			-0.16807
30	9802	MC4	0.42132			0.42132
31	9804	MC4	-0.49607			-0.49607
32	9803	MC4	-1.76786			-1.76786
33	9798	MC4	0.11274			0.11274
34	9800	EBSR4	1.21402			1.21402
35	9266	MC4	-1.04177			-1.04177
36	9272	MC4	-0.01006			-0.01006
37	9264	MC4	1.01395			1.01395
38	9262	MC4	1.18676			1.18676

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9267	HT	-0.59143			-0.59143
40	9263	MC4	-0.51352			-0.51352
41	9268	HT	2.08178			2.08178
42	10280	MC4	0.14517			0.14517
43	9872	HT	1.87785			1.87785
44	9870	HT	-0.18734			-0.18734
45	9869	MC4	-1.28088			-1.28088
46	9889	MC4	0.04903			0.04903
47	9867	MC4	0.88477			0.88477
48	9865	MC4	0.65339			0.65339
49	9866	MC4	0.07074			0.07074
50	9138	MC4	-1.48057			-1.48057
51	9168	EBSR4	0.35054			0.35054
52	9153	MC4	-0.41536			-0.41536
53	9130	MC4	-0.45134			-0.45134
54	9135	MS6	0.52718			0.52718
55	9134	MC4	-0.24354			-0.24354
56	9169	EBSR4	1.4272			1.4272
57	9131	MC4	0.87684			0.87684
58	9108	ETC	-1.32061	1.24437		-0.03812
59	9109	ETC	-1.49476	0.62001		-0.43738

Appendix H.5—Spring 16 Operational Item Parameter Estimates — Grade 7 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13400_e	ER	-0.89746	3.17022	4.70894	2.32723
2	13400_o	ER	-1.10875	2.81094	4.67186	2.12468
3	13400_c	ER	-1.93478	0.24135		-0.84672
4	13401_e	ER	-1.08712	3.58298	4.84348	2.44645
5	13401_o	ER	-2.01399	2.99796	5.04182	2.0086
6	13401_c	ER	-2.93019	-0.87734		-1.90377
7	13402_e	ER	-1.37884	3.89633	4.60481	2.3741
8	13402_o	ER	-1.49334	3.23935	5.1219	2.2893
9	13402_c	ER	-2.56472	-0.19658		-1.38065
10	13403_e	ER	0.48951	3.69526	5.16521	3.11666
11	13403_o	ER	-0.54572	3.33949	5.13751	2.64376
12	13403_c	ER	-2.82848	-0.61627		-1.72238
13	13405_e	ER	0.02255	3.38944	4.52755	2.64651
14	13405_o	ER	-0.88948	3.04194	4.92252	2.35833
15	13405_c	ER	-2.87726	-0.07362		-1.47544
16	13406_e	ER	-0.83081	4.35574	5.68822	3.07105
17	13406_o	ER	-1.24126	3.45562	5.9479	2.72075
18	13406_c	ER	-2.53963	0.10313		-1.21825
19	9103	ETC	0.33393			0.33393
20	9104	ETC	-1.21947	0.61781		-0.30083
21	9105	ETC	-0.14173			-0.14173
22	9147	MC4	-0.54315			-0.54315
23	9146	MC4	-1.19732			-1.19732
24	9152	MC4	-0.24903			-0.24903
25	9128	MS6	1.28632			1.28632
26	9610	MC4	0.90991			0.90991
27	9711	MC4	0.29644			0.29644
28	9611	HT	0.44176			0.44176
29	9713	MC4	0.2301			0.2301
30	9614	MC4	-1.02121			-1.02121
31	10613	MS5	1.17647			1.17647
32	10695	MC4	-0.51732			-0.51732
33	9750	MS6	-0.17847			-0.17847
34	9709	MC4	-0.24287			-0.24287
35	9177	MC4	0.5476			0.5476
36	9220	MC4	-0.2262			-0.2262
37	9176	MS5	1.43059			1.43059
38	10619	MC4	-0.10558			-0.10558

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9982	EBSR5	0.37426			0.37426
40	9962	MC4	0.59182			0.59182
41	10623	MC4	0.30353			0.30353
42	10621	EBSR4	1.03655			1.03655
43	9189	MC4	1.13439			1.13439
44	9778	MC4	0.42645			0.42645
45	9777	MC4	0.77082			0.77082
46	9779	HT	0.32439			0.32439
47	9781	MC4	0.22608			0.22608
48	9817	MS5	-0.10239			-0.10239
49	9811	MC4	-0.37913			-0.37913
50	9785	MS6	1.35871			1.35871
51	9786	MS5	1.03456			1.03456
52	9787	EBSR4	0.58622			0.58622
53	10274	MC4	1.70658			1.70658
54	9793	HT	1.62756			1.62756
55	9791	MC4	0.26709			0.26709
56	9789	MC4	-0.24857			-0.24857
57	10656	ETC	0.00688			0.00688
58	10657	ETC	-1.30701	-0.11629		-0.71165
59	10658	ETC	-0.75798	1.82429		0.53316

Appendix H.6—Spring 16 Operational Item Parameter Estimates — Grade 8 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13437_e	ER	-1.09959	1.99003	3.43357	1.44134
2	13437_o	ER	-1.21497	1.49481	3.52606	1.26863
3	13437_c	ER	-2.194	-0.55737		-1.37569
4	13438_e	ER	-1.30982	1.60181	3.36366	1.21855
5	13438_o	ER	-1.25185	1.38879	3.19409	1.11034
6	13438_c	ER	-2.48946	-0.96722		-1.72834
7	13439_e	ER	-1.77948	1.72887	2.90301	0.9508
8	13439_o	ER	-1.71597	1.09631	3.03584	0.80539
9	13439_c	ER	-2.40518	-1.1545		-1.77984
10	13452_e	ER	-1.67899	1.57768	3.33549	1.07806
11	13452_o	ER	-1.89107	1.08197	3.42027	0.87039
12	13452_c	ER	-2.54121	-1.11311		-1.82716
13	13453_e	ER	-1.06493	1.5231	3.24717	1.23511
14	13453_o	ER	-1.23977	0.91325	3.22926	0.96758
15	13453_c	ER	-1.99191	-0.7686		-1.38026
16	13454_e	ER	-1.16469	1.93019	3.31328	1.35959
17	13454_o	ER	-1.30159	1.37223	3.31286	1.12783
18	13454_C	ER	-2.12549	-0.79592		-1.46071
19	9116	MC4	-0.28257			-0.28257
20	9170	EBSR6	0.50345			0.50345
21	9115	HT	0.74599			0.74599
22	9240	MC4	-0.35422			-0.35422
23	9172	EBSR4	0.50692			0.50692
24	9113	MC4	-0.18363			-0.18363
25	9171	EBSR4	1.84183			1.84183
26	9255	HT	0.69969			0.69969
27	9259	MC4	0.64706			0.64706
28	9256	MC4	-0.24916			-0.24916
29	9253	MC4	-0.62194			-0.62194
30	9254	MC4	-0.32173			-0.32173
31	9017	MC4	-0.09267			-0.09267
32	9019	MC4	0.04892			0.04892
33	9014	MC4	-0.9944			-0.9944
34	9230	MC4	-1.15492			-1.15492
35	9018	MC4	0.80965			0.80965
36	9046	MC4	-0.82499			-0.82499
37	9015	MS5	1.20929			1.20929
38	9076	ETC	-0.64308			-0.64308

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9077	ETC	-1.7963	0.32962		-0.73334
40	9078	ETC	-0.93582	0.80924		-0.06329
41	10639	HT	2.91528			2.91528
42	9881	MC4	0.56406			0.56406
43	9880	MC4	-0.04633			-0.04633
44	9883	MC4	-1.4832			-1.4832
45	9885	MC4	-0.67563			-0.67563
46	10626	HT	1.26297			1.26297
47	9026	MC4	0.5053			0.5053
48	9022	MC4	0.99092			0.99092
49	9024	MC4	-0.53744			-0.53744
50	9023	MC4	0.95894			0.95894
51	9028	MC4	0.17704			0.17704
52	9020	MS5	1.09383			1.09383
53	9728	ETC	-1.65694	-0.20084		-0.92889
54	9729	ETC	-1.73003	0.26659		-0.73172
55	8965	MC4	-1.58511			-1.58511
56	8973	HT	-0.21957			-0.21957
57	9044	MC4	0.89154			0.89154
58	8967	MC4	-0.76092			-0.76092
59	8971	HT	1.15521			1.15521

Appendix H.7—Spring 16 Operational Item Parameter Estimates — Grade 9 ELA

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		13554_e	ER	-1.58337	2.72154	4.17335	1.77051
2		13554_o	ER	-1.49156	1.80507	4.60922	1.64091
3		13554_c	ER	-2.17867	-1.14465		-1.66166
4		13555_e	ER	-1.46735	2.51036	4.0225	1.6885
5		13555_o	ER	-1.7157	1.64362	3.78329	1.23707
6		13555_c	ER	-2.4999	-1.03399		-1.76695
7		13557_e	ER	-1.48371	2.29755	4.26558	1.69314
8		13557_o	ER	-1.33444	1.51754	4.19585	1.45965
9		13557_c	ER	-2.4354	-0.8662		-1.6508
10		13565_e	ER	-1.41497	1.96395	3.80007	1.44968
11		13565_o	ER	-1.69345	1.40948	3.82485	1.18029
12		13565_c	ER	-2.20831	-0.65189		-1.4301
13		13566_e	ER	-1.80017	2.23734	4.14528	1.52748
14		13566_o	ER	-1.66128	1.36788	4.09913	1.26858
15		13566_c	ER	-2.39075	-0.81852		-1.60464
16		13556_e	ER	-1.38759	2.43865	3.83392	1.62833
17		13556_o	ER	-1.76241	1.77841	3.89649	1.30416
18		13556_c	ER	-2.38602	-1.37075		-1.87839
19		10596	MC4	-1.63193			-1.63193
20		10595	MC4	-1.00777			-1.00777
21		10597	MC4	-1.30579			-1.30579
22		10603	MS6	0.37171			0.37171
23		9064	HT	1.21086			1.21086
24		9060	HT	1.2773			1.2773
25		9066	MC4	0.37042			0.37042
26		9065	MC4	0.29345			0.29345
27		9058	MC4	0.35867			0.35867
28		9063	MC4	-0.41625			-0.41625
29		9069	HT	1.02115			1.02115
30		1444	MC4	-0.4155			-0.4155
31		1446	MS4	-0.02528			-0.02528
32		9050	HT	1.37944			1.37944
33		9047	MS5	1.04482			1.04482
34		9048	MC4	-0.38622			-0.38622
35		9051	EBSR4	0.20427			0.20427
36		9053	MC4	-0.83869			-0.83869
37		11097	MC4	-0.45154			-0.45154
38		11098	MC4	0.4152			0.4152

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		9032	MC4	0.81518			0.81518
40		9040	MC4	1.00112			1.00112
41		9038	MC4	0.55427			0.55427
42		9033	MC4	-0.15469			-0.15469
43		9034	HT	1.18483			1.18483
44		8987	MC4	-0.57348			-0.57348
45		8989	MC4	-1.00653			-1.00653
46		9003	MC4	0.30221			0.30221
47		9005	HT	0.43076			0.43076
48		9043	EBSR4	0.37237			0.37237
49		8990	MC4	0.50835			0.50835
50		9004	MC4	0.0071			0.0071
51		9042	GI	0.59316			0.59316
52		9006	MC4	0.07488			0.07488
53		9007	MC4	-0.14351			-0.14351
54		9041	EBSR4	1.83768			1.83768
55		9012	MC4	0.3965			0.3965
56		8948	ETC	0.18506			0.18506
57		8951	ETC	-0.89183	1.4474		0.27779
58		9734	ETC	-0.38567			-0.38567
59		9735	ETC	-1.01417	1.57881		0.28232
60		9736	ETC	-0.28792			-0.28792
61		9737	ETC	-1.41093			-1.41093

Appendix H.8—Spring 16 Operational Item Parameter Estimates — Grade 10 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13638_e	ER	-1.78924	1.61688	3.39067	1.07277
2	13638_o	ER	-1.97581	0.93168	3.19227	0.71605
3	13638_c	ER	-3.08122	-1.11156		-2.09639
4	13635_e	ER	-1.64858	1.09969	3.31762	0.92291
5	13635_o	ER	-2.00564	0.76273	3.12769	0.62826
6	13635_c	ER	-3.09387	-1.09958		-2.09673
7	13639_e	ER	-1.16138	1.59043	3.81864	1.4159
8	13639_o	ER	-1.47773	1.06526	3.81875	1.13543
9	13639_c	ER	-2.96755	-1.16419		-2.06587
10	13636_e	ER	-1.14411	1.51623	4.30412	1.55875
11	13636_o	ER	-1.25137	0.71571	3.49201	0.98545
12	13636_c	ER	-2.77526	-1.56572		-2.17049
13	13637_e	ER	-1.2983	2.05466	3.36557	1.37398
14	13637_o	ER	-1.89522	1.40099	3.39837	0.96805
15	13637_c	ER	-2.97674	-0.96661		-1.97168
16	13640_e	ER	-1.90091	1.21808	2.96704	0.7614
17	13640_o	ER	-2.36052	0.51171	3.10403	0.41841
18	13640_c	ER	-3.03408	-1.28609		-2.16009
19	8785	MC4	-1.24117			-1.24117
20	8925	MC4	-0.35929			-0.35929
21	8788	MC4	-0.79543			-0.79543
22	8795	HT	1.52988			1.52988
23	8844	EBSR4	3.08365			3.08365
24	9623	MC4	-0.33285			-0.33285
25	9624	HT	0.96931			0.96931
26	9627	MC4	0.01553			0.01553
27	9626	MC4	0.39998			0.39998
28	9630	MC4	0.53771			0.53771
29	8815	MC4	0.55677			0.55677
30	8972	MC4	0.46067			0.46067
31	8816	MC4	0.64997			0.64997
32	8926	MC4	0.42232			0.42232
33	8821	MC4	0.18705			0.18705
34	8818	MC4	0.40891			0.40891
35	8819	MC4	0.62246			0.62246
36	9822	MC4	-0.54876			-0.54876
37	9824	MS5	0.82801			0.82801
38	9827	MC4	0.33355			0.33355

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9826	MC4	-0.43866			-0.43866
40	9821	MC4	0.03241			0.03241
41	9825	MC4	-0.52709			-0.52709
42	9887	MS5	1.98708			1.98708
43	9888	MC4	-0.02807			-0.02807
44	8760	ETC	0.86409			0.86409
45	8761	ETC	-0.1046			-0.1046
46	8762	ETC	0.57993			0.57993
47	8764	ETC	-0.66888	1.12054		0.22583
48	9831	MC4	-1.13658			-1.13658
49	9830	HT	1.81934			1.81934
50	9829	MC4	-0.02291			-0.02291
51	9837	MC4	-0.3871			-0.3871
52	9839	MC4	-0.068			-0.068
53	9840	MC4	0.4925			0.4925
54	9769	MC4	0.02052			0.02052
55	9773	MC4	0.8572			0.8572
56	9774	MC4	-0.07955			-0.07955
57	9772	MC4	-1.27366			-1.27366
58	9771	MC4	-0.44234			-0.44234
59	8757	ETC	-1.36906			-1.36906
60	8758	ETC	0.32886			0.32886
61	8759	ETC	-1.17573	-0.21852		-0.69713

Appendix H.9—Spring 16 Operational Item Parameter Estimates — Grade 11 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13723_e	ER	-1.9275	0.81378	2.56933	0.4852
2	13723_o	ER	-3.05795	0.1766	2.29503	-0.19544
3	13723_c	ER	-2.93586	-1.14637		-2.04112
4	13725_e	ER	-1.76303	0.57642	2.89886	0.57075
5	13725_o	ER	-2.23547	0.75545	2.49751	0.33916
6	13725_c	ER	-2.62307	-0.55033		-1.5867
7	13720_e	ER	-1.06366	1.77264	3.26664	1.32521
8	13720_o	ER	-2.24515	1.13742	3.0911	0.66112
9	13720_c	ER	-3.06998	-0.79431		-1.93215
10	13724_e	ER	-1.89212	0.54532	3.32819	0.66046
11	13724_o	ER	-2.63027	0.506	3.14192	0.33922
12	13724_c	ER	-2.90567	-0.96473		-1.9352
13	13721_e	ER	-2.24616	0.86028	3.32082	0.64498
14	13721_o	ER	-2.99104	0.50579	3.23517	0.24997
15	13721_c	ER	-2.98695	-0.98931		-1.98813
16	13722_e	ER	-1.92023	0.91135	3.08764	0.69292
17	13722_o	ER	-2.77119	0.40624	2.72972	0.12159
18	13722_C	ER	-2.88156	-1.06637		-1.97397
19	9853	HT	-0.7513			-0.7513
20	9856	MC4	-0.13181			-0.13181
21	9860	MC4	-0.3386			-0.3386
22	9858	MC4	0.99024			0.99024
23	9852	MS6	-0.84321			-0.84321
24	8862	EBSR4	0.76635			0.76635
25	8861	MC4	0.25879			0.25879
26	8865	HT	0.57228			0.57228
27	8867	MC4	0.18188			0.18188
28	8871	MC4	0.03863			0.03863
29	8805	HT	0.28483			0.28483
30	8807	MC4	0.03819			0.03819
31	8808	MC4	-0.43028			-0.43028
32	8809	HT	0.11628			0.11628
33	8846	EBSR4	0.2237			0.2237
34	8753	ETC	-1.29646			-1.29646
35	8754	ETC	-1.27686			-1.27686
36	8755	ETC	-0.48937	1.65996		0.5853
37	8797	MC4	-0.35939			-0.35939
38	8798	MC4	0.17292			0.17292

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	8796	MC4	0.05932			0.05932
40	8800	MC4	-0.77328			-0.77328
41	8799	MC4	-0.52283			-0.52283
42	8880	MC4	-0.02503			-0.02503
43	8881	MC4	-0.37143			-0.37143
44	8879	MS6	0.52607			0.52607
45	8884	MC4	0.26397			0.26397
46	8794	HT	0.60324			0.60324
47	8783	MC4	-0.68116			-0.68116
48	8792	MC4	-0.21395			-0.21395
49	8784	MC4	-0.25728			-0.25728
50	8781	MC4	0.34351			0.34351
51	8791	MS5	1.31404			1.31404
52	421	MC4	-0.19236			-0.19236
53	422	MS4	1.07426			1.07426
54	8856	EBSR4	0.53849			0.53849
55	8834	MC4	-0.09598			-0.09598
56	8837	MS5	1.07611			1.07611
57	8841	MC4	0.06178			0.06178
58	8843	MS5	0.82014			0.82014
59	8769	ETC	0.3946			0.3946
60	8770	ETC	-1.5716	0.25326		-0.65917
61	8771	ETC	-1.2835			-1.2835

Appendix H.10—Spring 16 Operational Item Parameter Estimates — Grade 3 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10453	MC4	-3.3943			-3.3943
2	10425	MC4	-1.53829			-1.53829
3	10411	MC4	-1.61799			-1.61799
4	10403	MC4	-0.58186			-0.58186
5	8463	MC4	-0.59241			-0.59241
6	9455	EQ	-0.07043			-0.07043
7	10416	MC4	0.31901			0.31901
8	10427	EQ	1.0682			1.0682
9	10448	EQ	1.09544			1.09544
10	10430	EQ	2.32239			2.32239
11	10685	MC4	1.69992			1.69992
12	10470	EQ	2.23998			2.23998
13	10399	EQ	0.88755			0.88755
14	10446	EQ	1.29275			1.29275
15	10398	EQ	0.4498			0.4498
16	10396	EQ	0.38434			0.38434
17	10671	EQ	0.39429			0.39429
18	10400	EQ	-0.25945			-0.25945
19	10434	EQ	-0.75061			-0.75061
20	8461	MC4	-0.60467			-0.60467
21	10395	EQ	-2.04152			-2.04152
22	10438	EQ	-2.34273			-2.34273
23	10443	MC4	-2.32838			-2.32838
24	10409	MC4	-1.87291			-1.87291
25	10384	EQ	-1.35609			-1.35609
26	10439	EQ	-1.01284			-1.01284
27	10436	MC4	-0.50493			-0.50493
28	10391	MC4	-0.4991			-0.4991
29	10466	EQ	0.826			0.826
30	8481	MC4	0.82314			0.82314
31	10433	EQ	1.33708			1.33708
32	10455	EQ	1.85302			1.85302
33	9464	EQ	1.3174			1.3174
34	10454	MI	2.28432			2.28432
35	9469	EQ	1.57178			1.57178
36	11120	EQ	1.1283			1.1283
37	10683	MC4	1.42686			1.42686
38	10392	EQ	0.62557			0.62557

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10687	MC4	0.89578			0.89578
40	10418	EQ	-0.2178			-0.2178
41	10450	GI	0.34171			0.34171
42	8483	MC4	-0.02135			-0.02135
43	10431	EQ	-1.08021			-1.08021
44	9460	EQ	-1.67509			-1.67509
45	10402	EQ	-2.02582			-2.02582

Appendix H.11—Spring 16 Operational Item Parameter Estimates — Grade 4 Mathematics

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		9456	GI	-3.07549			-3.07549
2		10766	MS5	-1.46597			-1.46597
3		10771	MS5	-0.98077			-0.98077
4		10826	GI	-1.49232	0.80594		-0.34319
5		10710	MS5	-0.44114			-0.44114
6		9474	EQ	-0.17779			-0.17779
7		8501	MC4	0.50338			0.50338
8		10739	MC4	0.68294			0.68294
9		9465	EQ	0.63804			0.63804
10		10782	MS5	1.46482			1.46482
11		10754	MS5	1.42943			1.42943
12		10780	EQ	2.24981	2.50292		2.37636
13		9475	GI	2.19462			2.19462
14		10728	EQ	1.29344			1.29344
15		10724	MS5	1.13173			1.13173
16		10715	EQ	0.74127			0.74127
17		10752	MC4	0.44252			0.44252
18		9502	EQ	0.32547			0.32547
19		10718	EQ	0.03231			0.03231
20		10827	EQ	-0.86309			-0.86309
21		10783	MC4	-0.58052			-0.58052
22		10731	EQ	-0.7754			-0.7754
23		8505	MC4	-1.77368			-1.77368
24		10772	MC4	-2.60398			-2.60398
25		10729	EQ	-0.97271			-0.97271
26		8497	MC4	-1.38394			-1.38394
27		10730	EQ	-0.92949			-0.92949
28		10750	EQ	-0.58904			-0.58904
29		10705	MC4	0.15075			0.15075
30		10744	EQ	-0.50014			-0.50014
31		9452	EQ	0.34683			0.34683
32		9467	GI	0.92347			0.92347
33		10716	EQ	1.10308			1.10308
34		9470	EQ	1.96056			1.96056
35		10727	EQ	2.15749			2.15749
36		10753	MS5	0.86014			0.86014
37		10779	EQ	0.34347			0.34347
38		10781	MC4	0.68116			0.68116

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		9471	EQ	1.32625			1.32625
40		10719	EQ	-0.03793			-0.03793
41		10763	EQ	-0.20366			-0.20366
42		10751	MC4	-0.3404			-0.3404
43		10708	MC4	-0.36231			-0.36231
44		8493	MC4	-0.80101			-0.80101
45		8499	MC4	-2.11291			-2.11291

Appendix H.12—Spring 16 Operational Item Parameter Estimates — Grade 5 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10789	EQ	-2.83917			-2.83917
2	10811	MC4	-1.11389			-1.11389
3	10829	EQ	-1.59476			-1.59476
4	10836	MC4	-0.75042			-0.75042
5	10799	EQ	-0.59379			-0.59379
6	9486	EQ	0.14711			0.14711
7	10803	EQ	-0.2183			-0.2183
8	10839	EQ	0.63393			0.63393
9	10808	EQ	0.95699			0.95699
10	10813	EQ	1.41929			1.41929
11	9476	EQ	1.35515			1.35515
12	10835	EQ	2.12229			2.12229
13	10832	MC4	-0.6078			-0.6078
14	10849	EQ	1.15146			1.15146
15	10788	MS5	0.2804			0.2804
16	10863	MC4	0.62779			0.62779
17	10824	MC4	-0.17851			-0.17851
18	10798	MS5	-0.17942			-0.17942
19	8525	MC4	-0.58567			-0.58567
20	10872	GI	-0.76667	-0.47954		-0.6231
21	8535	MC4	-0.97262			-0.97262
22	8539	MC4	-1.17271			-1.17271
23	9716	MI	-2.30144			-2.30144
24	10848	EQ	-0.81249			-0.81249
25	8521	MC4	-1.17672			-1.17672
26	10790	EQ	0.78506			0.78506
27	11107	MC4	-0.46606			-0.46606
28	10805	EQ	-0.48618			-0.48618
29	10850	EQ	0.19694			0.19694
30	10816	EQ	0.02499			0.02499
31	10851	EQ	0.92342			0.92342
32	10833	EQ	-0.99926			-0.99926
33	9485	GI	1.76276			1.76276
34	10868	EQ	3.6372			3.6372
35	10817	TI	3.1767	1.86902		2.52286
36	10840	EQ	0.96138			0.96138
37	10820	MC4	1.62329			1.62329
38	9487	EQ	1.1116			1.1116

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10796	EQ	0.28237			0.28237
40	10869	MS6	0.21386			0.21386
41	10800	EQ	1.11572			1.11572
42	8527	MC4	-0.68748			-0.68748
43	10791	EQ	0.89984			0.89984
44	10875	MC4	-0.69325			-0.69325
45	10795	EQ	-1.39678			-1.39678

Appendix H.13—Spring 16 Operational Item Parameter Estimates — Grade 6 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10067	EQ	-3.72557			-3.72557
2	9491	GI	-1.70216			-1.70216
3	9492	EQ	-1.41594			-1.41594
4	10053	EQ	-0.7795			-0.7795
5	8567	MC4	-0.65504			-0.65504
6	10051	EQ	-0.35771			-0.35771
7	10120	EQ	0.47249			0.47249
8	10076	TI	0.82913			0.82913
9	10083	MS5	0.91604			0.91604
10	10096	EQ	0.93422			0.93422
11	9512	EQ	1.37691			1.37691
12	9718	MI	2.34031			2.34031
13	10151	GI	1.98277			1.98277
14	10117	MC4	0.89436			0.89436
15	10143	EQ	0.87374			0.87374
16	10095	TI	0.97298			0.97298
17	10071	MC4	0.56468			0.56468
18	10093	MC4	0.03634			0.03634
19	10064	EQ	-0.12482			-0.12482
20	10052	MC4	0.15283			0.15283
21	10115	EQ	-0.21997			-0.21997
22	10113	MC4	-1.43848			-1.43848
23	10137	EQ	-2.35826			-2.35826
24	8549	MC4	-2.53259			-2.53259
25	10106	EQ	-1.86826			-1.86826
26	9719	MI	-1.29034			-1.29034
27	10108	MC4	-0.6664			-0.6664
28	10057	EQ	-1.19929			-1.19929
29	9496	EQ	0.22756			0.22756
30	10049	EQ	-0.09381			-0.09381
31	10048	MC4	0.40685			0.40685
32	10078	EQ	1.05127			1.05127
33	10060	EQ	1.20803			1.20803
34	9513	GI	1.29516			1.29516
35	9498	EQ	2.73279			2.73279
36	10139	MS6	2.24036			2.24036
37	10111	MS5	1.60177			1.60177
38	10070	EQ	0.2958			0.2958

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10094	EQ	1.03964			1.03964
40	10079	EQ	0.82051			0.82051
41	10088	EQ	0.63477			0.63477
42	10150	EQ	-0.04156			-0.04156
43	8555	MC4	-0.17599			-0.17599
44	10148	MC4	-0.4154			-0.4154
45	10103	EQ	-0.93491			-0.93491
46	10107	EQ	-1.82823			-1.82823
47	10129	MC4	-2.30001			-2.30001

Appendix H.14—Spring 16 Operational Item Parameter Estimates — Grade 7 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	8698	MC4	-2.05133			-2.05133
2	10290	MC4	-0.68728			-0.68728
3	10301	EQ	-0.54092			-0.54092
4	8597	MC4	0.07583			0.07583
5	9520	GI	0.27072			0.27072
6	10294	EQ	0.11229			0.11229
7	10701	EQ	1.03009			1.03009
8	10298	EQ	0.12637			0.12637
9	10351	MC4	1.341			1.341
10	10302	EQ	1.16441			1.16441
11	10317	EQ	1.81273			1.81273
12	10347	EQ	1.81767			1.81767
13	10340	GI	1.98187			1.98187
14	10344	MS5	1.05834			1.05834
15	10341	GI	2.17195			2.17195
16	9514	GI	0.89724			0.89724
17	10374	EQ	0.52966			0.52966
18	8609	MC4	0.36499			0.36499
19	10366	EQ	0.38042	0.32945		0.35494
20	9516	EQ	0.28886			0.28886
21	10322	EQ	-0.24088			-0.24088
22	8593	MC4	-0.19898			-0.19898
23	8603	MC4	-0.44189			-0.44189
24	10378	MC4	-2.47548			-2.47548
25	8613	MC4	-1.05036			-1.05036
26	13863	MC4	-1.02788			-1.02788
27	10303	MC4	-1.12478			-1.12478
28	10288	EQ	-0.97446			-0.97446
29	13865	MC4	-0.8631			-0.8631
30	10362	MI	-0.5352			-0.5352
31	13856	MC4	-0.56928			-0.56928
32	10349	MC4	-0.06195			-0.06195
33	13867	MC4	1.17205			1.17205
34	13857	MC4	0.53956			0.53956
35	10375	MC4	1.72141			1.72141
36	10352	MC4	1.07916			1.07916
37	10309	MI	0.97888			0.97888
38	10369	MC4	0.25616			0.25616

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10331	EQ	0.38423			0.38423
40	9508	EQ	1.88124			1.88124
41	13860	MC4	-0.51315			-0.51315
42	10318	MC4	-0.85168			-0.85168
43	10299	EQ	-0.49012			-0.49012
44	10313	MC4	-0.82795			-0.82795
45	10379	GI	0.47078			0.47078
46	10371	MC4	-1.15822			-1.15822
47	13859	MC4	-1.02675			-1.02675

Appendix H.15—Spring 16 Operational Item Parameter Estimates — Grade 8 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10488	TI	-2.19043			-2.19043
2	10530	MC4	-0.83532			-0.83532
3	10567	EQ	-0.90128			-0.90128
4	10561	MC4	-0.22329			-0.22329
5	9532	EQ	0.73485			0.73485
6	10532	EQ	0.16579			0.16579
7	9521	EQ	1.11345			1.11345
8	10541	EQ	0.75865			0.75865
9	10585	EQ	1.6013			1.6013
10	10570	MI	1.10187			1.10187
11	9522	EQ	2.50544			2.50544
12	10579	EQ	3.02045			3.02045
13	9518	EQ	2.17755			2.17755
14	10523	EQ	2.68961			2.68961
15	10542	EQ	1.16308			1.16308
16	10538	EQ	1.41298			1.41298
17	10564	EQ	0.40128			0.40128
18	10580	EQ	0.96886			0.96886
19	9519	GI	-0.04329			-0.04329
20	10587	MC4	0.43524			0.43524
21	10513	MC4	-0.21755			-0.21755
22	10483	EQ	-0.67181			-0.67181
23	8623	MC4	-1.63504			-1.63504
24	10557	MC4	-3.16617			-3.16617
25	10562	MC4	-0.80153			-0.80153
26	10507	MC4	-1.82784			-1.82784
27	10498	MC4	-0.91777			-0.91777
28	10588	MC4	-1.06474			-1.06474
29	10554	MC4	-1.02849			-1.02849
30	10496	GI	-0.34275			-0.34275
31	10548	GI	0.35118			0.35118
32	9527	GI	-0.043			-0.043
33	10494	MS5	-0.08244			-0.08244
34	8651	MC4	-0.12423			-0.12423
35	10518	TI	1.44965			1.44965
36	9525	EQ	0.25872	0.32204		0.29038
37	10510	MC4	0.71001			0.71001
38	10487	MC4	-0.12623			-0.12623

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10527	EQ	0.11829			0.11829
40	10534	GI	0.04773			0.04773
41	10574	MS6	-0.26598			-0.26598
42	8635	MC4	-0.83239			-0.83239
43	10520	GI	-0.95493			-0.95493
44	8631	MC4	-1.22684			-1.22684
45	10525	MC4	-1.64422			-1.64422
46	10581	MC4	-1.92813			-1.92813
47	10528	EQ	-2.45491			-2.45491

Appendix H.16—Spring 16 Operational Item Parameter Estimates — Algebra I

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		10882	EQ	0.8083			0.8083
2		10888	EQ	1.89687			1.89687
3		10960	MC4	0.60667			0.60667
4		10951	MC4	0.71118			0.71118
5		11013	MS5	0.09622			0.09622
6		9530	TI	0.69364			0.69364
7		10972	EQ	-0.09826			-0.09826
8		11045	MC4	-0.09418			-0.09418
9		10895	EQ	-0.05275			-0.05275
10		10977	MC4	-0.42252			-0.42252
11		10953	MC4	-0.75521			-0.75521
12		10934	EQ	-0.94776			-0.94776
13		9707	GI	-2.17345			-2.17345
14		10887	MC4	-1.07459			-1.07459
15		10907	MC4	-1.12152			-1.12152
16		9705	GI	-0.96679			-0.96679
17		10974	MC4	-0.58968			-0.58968
18		11058	MC4	0.27251			0.27251
19		10889	MC4	-0.16663			-0.16663
20		10966	MC4	0.15284			0.15284
21		10943	MC4	0.46515			0.46515
22		10990	MC4	0.65596			0.65596
23		10978	MS5	1.8257			1.8257
24		11044	EQ	1.12987			1.12987
25		9546	EQ	0.79976			0.79976
26		9536	GI	1.36497			1.36497
27		11011	EQ	0.32291			0.32291
28		10896	MC4	0.61086			0.61086
29		11004	MC4	0.25768			0.25768
30		10942	MC4	0.09573			0.09573
31		10935	MC4	-0.12131			-0.12131
32		10945	MC4	-0.65134			-0.65134
33		9543	GI	-0.70978	-0.39294		-0.55136
34		10993	MC4	-1.08396			-1.08396
35		10905	MC4	-0.97781			-0.97781
36		10973	MC4	-1.36424			-1.36424
37		11052	MC4	-1.26248			-1.26248
38		10963	MC4	-0.87904			-0.87904

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		10877	MC4	-0.39859			-0.39859
40		9535	EQ	-0.04011			-0.04011
41		10880	MC4	1.31529			1.31529
42		10897	MC4	0.32317			0.32317
43		10906	MC4	0.18842			0.18842
44		9533	TI	0.7707			0.7707
45		10988	EQ	0.41304			0.41304
46		10941	MC4	1.52007			1.52007
47		9531	EQ	0.97351			0.97351

Appendix H.17—Spring 16 Operational Item Parameter Estimates — Geometry

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	11037	MC4	-3.17485			-3.17485
2	10919	MC4	-1.19939			-1.19939
3	11018	MC4	-0.61172			-0.61172
4	11072	MC4	-0.56959			-0.56959
5	11089	MS6	-0.7697			-0.7697
6	10924	EQ	-0.20101			-0.20101
7	9556	EQ	-0.21949			-0.21949
8	10912	MC4	0.70057			0.70057
9	10926	EQ	0.4718			0.4718
10	9554	HT	0.94194			0.94194
11	9722	MI	1.01157			1.01157
12	9592	GI	1.50932			1.50932
13	9575	EQ	2.83952			2.83952
14	11017	HT	0.72651			0.72651
15	11074	MS5	0.57924			0.57924
16	9581	HT	0.20433			0.20433
17	11008	MC4	0.53528			0.53528
18	11007	EQ	-0.74951			-0.74951
19	9560	HT	0.32402			0.32402
20	11068	MC4	-0.64576			-0.64576
21	11035	MC4	-0.58779			-0.58779
22	10986	MC4	-1.15355			-1.15355
23	11061	MC4	-1.23202			-1.23202
24	10921	MC4	-1.82488			-1.82488
25	11086	MC4	-1.06633			-1.06633
26	11033	MC4	-1.06679			-1.06679
27	11016	EQ	-0.8117			-0.8117
28	11040	EQ	-0.73442			-0.73442
29	11078	MC4	-0.56802			-0.56802
30	11029	MC4	-0.13274			-0.13274
31	9564	EQ	0.43952			0.43952
32	11085	HT	0.56151			0.56151
33	11032	HT	0.86558			0.86558
34	11065	EQ	1.01365			1.01365
35	10998	EQ	3.68979			3.68979
36	9551	HT	2.06285			2.06285
37	11081	MS5	1.69723			1.69723
38	11092	MS6	0.93109			0.93109

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	11060	MC4	0.47251			0.47251
40	10930	EQ	0.73492			0.73492
41	11034	MC4	-0.04112			-0.04112
42	11114	HT	-0.12484			-0.12484
43	11036	MC4	-0.31518			-0.31518
44	11026	HT	-0.76141			-0.76141
45	9547	EQ	-1.26142			-1.26142
46	10910	MC4	-0.7847			-0.7847
47	11015	EQ	-1.64967			-1.64967

Appendix H.18—Spring 16 Operational Item Parameter Estimates — Algebra II

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10215	MC4	-2.47281			-2.47281
2	10175	MC4	-1.28953			-1.28953
3	10210	MC4	-1.79638			-1.79638
4	10236	MC4	-0.89747			-0.89747
5	11121	MC4	-1.0416			-1.0416
6	9570	EQ	-0.10457			-0.10457
7	10245	EQ	0.65995			0.65995
8	9578	EQ	0.25925			0.25925
9	10164	EQ	0.69032			0.69032
10	9577	EQ	1.22957			1.22957
11	10182	EQ	1.37683			1.37683
12	10249	EQ	1.06528			1.06528
13	10256	GI	1.04876			1.04876
14	9548	EQ	2.2849			2.2849
15	10255	EQ	1.17865			1.17865
16	9567	EQ	0.98222			0.98222
17	10261	EQ	0.50767			0.50767
18	9549	EQ	0.03224			0.03224
19	10187	MC4	-0.17866			-0.17866
20	10233	EQ	-0.81277			-0.81277
21	10160	MC4	-0.65505			-0.65505
22	10192	MC4	-1.21466			-1.21466
23	10214	MC4	-1.65318			-1.65318
24	10206	MC4	-2.46401			-2.46401
25	10200	MC4	-1.46483			-1.46483
26	10204	HT	-1.37366			-1.37366
27	10203	MC4	-1.02449			-1.02449
28	10177	MC4	-0.56882			-0.56882
29	10259	MC4	-0.09362			-0.09362
30	10240	MS5	0.00695			0.00695
31	9568	EQ	0.85211			0.85211
32	10220	MC4	0.5445			0.5445
33	10168	MS5	1.69076			1.69076
34	10176	MS5	0.82007			0.82007
35	9589	EQ	2.2194			2.2194
36	9591	EQ	1.9907			1.9907
37	10180	EQ	2.12488			2.12488
38	10199	HT	0.94861			0.94861

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10223	EQ	0.48221			0.48221
40	10228	MI	0.55259			0.55259
41	9573	EQ	0.31592			0.31592
42	10230	MC4	-0.3605			-0.3605
43	9580	EQ	-0.05533			-0.05533
44	10209	MC4	-0.65893			-0.65893
45	10243	MC4	-0.46516			-0.46516
46	10217	MC4	-1.30014			-1.30014
47	10237	MC4	-1.70996			-1.70996

Appendix I.1 – Standard Errors of Measurement at Performance Level Cuts Spring 2016 – ELA

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 ELA	10.35	9.46	10.15	12.50	10.46
Grade 4 ELA	10.68	10.00	10.41	13.09	10.82
Grade 5 ELA	10.67	10.00	10.50	12.93	10.80
Grade 6 ELA	10.52	9.62	10.34	13.22	10.47
Grade 7 ELA	10.39	9.00	10.00	12.97	10.16
Grade 8 ELA	10.01	9.05	10.25	12.61	10.08
Grade 9 ELA	10.22	9.00	9.91	12.31	10.04
Grade 10 ELA	9.74	9.00	9.68	11.86	9.84
Grade 11 ELA	9.93	9.00	9.46	12.00	9.94

Appendix I.2 – Standard Errors of Measurement at Performance Level Cuts Spring 2016– Mathematics

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 Math	11.97	10.74	11.89	17.34	12.70
Grade 4 Math	11.53	10.00	11.07	15.54	11.49
Grade 5 Math	12.32	10.00	10.68	13.72	11.48
Grade 6 Math	12.50	10.00	10.42	13.83	11.67
Grade 7 Math	11.11	10.00	10.18	13.31	10.90
Grade 8 Math	11.81	10.00	10.20	12.68	11.24
Algebra I	11.39	9.55	9.55	13.21	10.76
Geometry	13.65	10.39	10.00	12.60	11.90
Algebra II	13.74	11.11	10.00	11.27	12.20

Appendix J.1 Writing Prompt Rater Agreement Report - Grade 3 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,155	4,155	0.9	16.2	65.8	16.1	0.9	0	0
	Evidence/Elaboration	4,155	4,155	1	16.5	65.1	16.3	1	0	0
	Conventions	4,155	4,155	1.1	14.5	68.7	14.6	1.1	0	0
2	Purpose/Organization	4,176	4,176	1.4	18.2	60.6	18.3	1.4	0	0
	Evidence/Elaboration	4,176	4,176	0.9	17.4	63.5	17.5	0.9	0	0
	Conventions	4,176	4,176	1.9	15.3	65.5	15.3	1.9	0	0
3	Purpose/Organization	4,153	4,153	0.7	17.3	64	17.2	0.7	0	0
	Evidence/Elaboration	4,153	4,153	0.7	16.5	65.7	16.4	0.7	0	0
	Conventions	4,153	4,153	1	13.7	70.8	13.6	1	0	0
4	Purpose/Organization	4,186	4,186	0.9	17.3	63.7	17.3	0.9	0	0
	Evidence/Elaboration	4,186	4,186	1.4	18.4	60.3	18.4	1.4	0	0
	Conventions	4,186	4,186	0.5	14	71.1	14	0.5	0	0
5	Purpose/Organization	4,176	4,176	1.1	17.1	63.4	17.2	1.1	0	0
	Evidence/Elaboration	4,176	4,176	1.1	17.6	62.5	17.6	1.1	0	0
	Conventions	4,176	4,176	0.8	13.5	71.4	13.5	0.7	0	0
6	Purpose/Organization	4,175	4,175	1.2	16.8	64	16.9	1.2	0	0
	Evidence/Elaboration	4,175	4,175	1.7	17.5	61.5	17.6	1.7	0	0
	Conventions	4,175	4,175	0.7	14.2	70.2	14.2	0.7	0	0
Paper Essay	Purpose/Organization	28,556	5,526	0.4	3.4	92.5	3.4	0.4	0	0
	Evidence/Elaboration	28,556	5,526	0.3	3.6	92.2	3.6	0.3	0	0
	Conventions	28,556	5,526	0.2	2.8	94	2.8	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.2 Writing Prompt Rater Agreement Report - Grade 4 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,190	4,190	0.3	12.7	74	12.7	0.3	0	0
	Evidence/Elaboration	4,190	4,190	0.1	11.3	77.2	11.3	0.1	0	0
	Conventions	4,190	4,190	1	15.8	66.4	15.8	1	0	0
2	Purpose/Organization	4,189	4,189	0.4	14.2	70.9	14.2	0.4	0	0
	Evidence/Elaboration	4,189	4,189	0.2	12.3	75.1	12.3	0.2	0	0
	Conventions	4,189	4,189	0.3	17.8	63.7	17.8	0.3	0	0
3	Purpose/Organization	4,188	4,188	0.3	14	71.5	14	0.3	0	0
	Evidence/Elaboration	4,188	4,188	0.2	13	73.5	13	0.2	0	0
	Conventions	4,188	4,188	0.3	16.6	66.2	16.6	0.3	0	0
4	Purpose/Organization	4,192	4,192	0.4	13.8	71.8	13.8	0.4	0	0
	Evidence/Elaboration	4,192	4,192	0.2	11.5	76.8	11.5	0.2	0	0
	Conventions	4,192	4,192	0.2	16.2	67.1	16.2	0.2	0	0
5	Purpose/Organization	4,192	4,192	0.4	14.1	71.1	14.1	0.4	0	0
	Evidence/Elaboration	4,192	4,192	0.3	11.9	75.7	11.9	0.3	0	0
	Conventions	4,192	4,192	0.5	17.3	64.2	17.3	0.5	0	0
6	Purpose/Organization	4,182	4,182	0.1	12.7	74.4	12.7	0.1	0	0
	Evidence/Elaboration	4,182	4,182	0.2	11	77.5	11	0.2	0	0
	Conventions	4,182	4,182	0.3	16.3	66.8	16.3	0.3	0	0
Paper Essay	Purpose/Organization	28,026	5,098	0.1	11.7	76.3	11.7	0.1	0	0
	Evidence/Elaboration	28,026	5,098	0.2	11	77.6	11	0.2	0	0
	Conventions	28,026	5,098	0.2	12.6	74.4	12.6	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.3 Writing Prompt Rater Agreement Report - Grade 5 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,186	4,186	0.2	14.5	70.5	14.5	0.2	0	0
	Evidence/Elaboration	4,186	4,186	0.2	14	71.5	14	0.2	0	0
	Conventions	4,186	4,186	0.2	13.6	72.3	13.6	0.2	0	0
2	Purpose/Organization	4,198	4,198	0.1	14.6	70.5	14.6	0.1	0	0
	Evidence/Elaboration	4,198	4,198	0.2	14.4	70.8	14.4	0.2	0	0
	Conventions	4,198	4,198	0.2	12.8	74.0	12.8	0.2	0	0
3	Purpose/Organization	4,190	4,190	0.2	13.9	71.8	13.9	0.2	0	0
	Evidence/Elaboration	4,190	4,190	0.3	13.4	72.6	13.4	0.3	0	0
	Conventions	4,190	4,190	0.2	13.4	72.7	13.4	0.2	0	0
4	Purpose/Organization	4,196	4,196	0.3	14.8	69.9	14.8	0.3	0	0
	Evidence/Elaboration	4,196	4,196	0.3	15.1	69.1	15.2	0.3	0	0
	Conventions	4,196	4,196	0.2	13.6	72.3	13.6	0.2	0	0
5	Purpose/Organization	4,192	4,192	0.2	14.2	71.1	14.2	0.2	0	0
	Evidence/Elaboration	4,192	4,192	0.5	16.8	65.3	16.8	0.5	0	0
	Conventions	4,192	4,192	0.5	12.9	73.2	12.9	0.5	0	0
6	Purpose/Organization	4,194	4,194	0.4	15.1	69.1	15.1	0.4	0	0
	Evidence/Elaboration	4,194	4,194	0.5	17	65.0	17	0.5	0	0
	Conventions	4,194	4,194	0.1	13.7	72.4	13.7	0.1	0	0
Paper Essay	Purpose/Organization	27,955	5,088	0.2	8.3	83.1	8.3	0.2	0	0
	Evidence/Elaboration	27,955	5,088	0.2	7.6	84.6	7.6	0.2	0	0
	Conventions	27,955	5,088	0.1	6.4	86.9	6.4	0.1	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.4 Writing Prompt Rater Agreement Report - Grade 6 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,190	4,190	1	17.8	62.4	17.8	1	0	0
	Evidence/Elaboration	4,190	4,190	0.8	15.7	66.9	15.7	0.8	0	0
	Conventions	4,190	4,190	0.5	16.4	66.2	16.4	0.5	0	0
2	Purpose/Organization	4,194	4,194	0.9	17.4	63.4	17.4	0.9	0	0
	Evidence/Elaboration	4,194	4,194	0.9	14.5	69.1	14.5	0.9	0	0
	Conventions	4,194	4,194	1.2	17.5	62.5	17.5	1.2	0	0
3	Purpose/Organization	4,186	4,186	1.4	18	61.2	18	1.4	0	0
	Evidence/Elaboration	4,186	4,186	1.5	16.1	64.8	16.1	1.5	0	0
	Conventions	4,186	4,186	0.6	15.2	68.2	15.2	0.6	0	0
4	Purpose/Organization	4,194	4,194	1.3	18	61.5	18	1.3	0	0
	Evidence/Elaboration	4,194	4,194	1.4	17.2	62.8	17.2	1.4	0	0
	Conventions	4,194	4,194	0.5	16.2	66.6	16.2	0.5	0	0
5	Purpose/Organization	4,190	4,190	0.6	16.7	65.4	16.7	0.6	0	0
	Evidence/Elaboration	4,190	4,190	0.2	14.8	69.8	14.8	0.2	0	0
	Conventions	4,190	4,190	0.7	16.1	66.4	16.1	0.7	0	0
6	Purpose/Organization	4,194	4,194	0.6	17.5	63.8	17.5	0.6	0	0
	Evidence/Elaboration	4,194	4,194	0.5	17.3	64.5	17.3	0.5	0	0
	Conventions	4,194	4,194	0.6	13.9	71.1	13.9	0.6	0	0
Paper Essay	Purpose/Organization	30,154	5,695	0.3	10.6	78.1	10.6	0.3	0	0
	Evidence/Elaboration	30,154	5,695	0.1	8.8	82.2	8.8	0.1	0	0
	Conventions	30,154	5,695	0.4	9.3	80.6	9.3	0.4	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.5 Writing Prompt Rater Agreement Report - Grade 7 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,164	4,164	0.6	16.7	65.4	16.7	0.6	0	0
	Evidence/Elaboration	4,164	4,164	0.6	16.5	65.7	16.5	0.6	0	0
	Conventions	4,164	4,164	0.2	14.9	69.7	14.9	0.2	0	0
2	Purpose/Organization	4,168	4,168	0.2	13.5	72.6	13.5	0.2	0	0
	Evidence/Elaboration	4,168	4,168	0.2	13.9	71.8	13.9	0.2	0	0
	Conventions	4,168	4,168	0.2	10.5	78.5	10.5	0.2	0	0
3	Purpose/Organization	4,176	4,176	0.1	15	69.8	15	0.1	0	0
	Evidence/Elaboration	4,176	4,176	0.2	13.3	73	13.3	0.2	0	0
	Conventions	4,176	4,176	0.2	15.3	68.9	15.3	0.2	0	0
4	Purpose/Organization	4,170	4,170	0.5	15.3	68.4	15.3	0.5	0	0
	Evidence/Elaboration	4,170	4,170	0.4	13.3	72.6	13.3	0.4	0	0
	Conventions	4,170	4,170	0.1	11.6	76.7	11.6	0.1	0	0
5	Purpose/Organization	4,186	4,186	0.1	13.4	73.1	13.4	0.1	0	0
	Evidence/Elaboration	4,186	4,186	0.2	12.7	74.2	12.7	0.2	0	0
	Conventions	4,186	4,186	0.1	12.4	75.1	12.4	0.1	0	0
6	Purpose/Organization	4,180	4,180	0.1	14.4	71	14.4	0.1	0	0
	Evidence/Elaboration	4,180	4,180	0.2	13.3	73	13.3	0.2	0	0
	Conventions	4,180	4,180	0.1	14.1	71.5	14.1	0.1	0	0
Paper Essay	Purpose/Organization	29,075	5,353	0.1	10.9	78.1	10.9	0.1	0	0
	Evidence/Elaboration	29,075	5,353	0.1	10.1	79.6	10.1	0.1	0	0
	Conventions	29,075	5,353	0.2	9.5	80.7	9.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.6 Writing Prompt Rater Agreement Report - Grade 8 ELA

				Rater Agreement						
	Dimension	Total Reads	Second Reads	Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	28,253	5,136	0.8	15.3	67.8	15.3	0.8	0	0.1
	Evidence/Elaboration	28,253	5,136	0.7	15.3	67.8	15.3	0.7	0	0.1
	Conventions	28,253	5,136	0.5	8.3	82.3	8.3	0.5	0	0.1

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.7 ELA Writing Prompt Rater Agreement Report - Grade 9 ELA

				Rater Agreement						
Dimension		Total Reads	Second Reads	Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	33,077	6,026	0.4	13.2	72.8	13.2	0.4	0	0
	Evidence/Elaboration	33,077	6,026	0.7	12.7	73.2	12.7	0.7	0	0
	Conventions	33,077	6,026	0.1	6.8	86.2	6.8	0.1	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.8 ELA Writing Prompt Rater Agreement Report - Grade 10 ELA

				Rater Agreement						
Dimension		Total Reads	Second Reads	Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	30,323	5,520	0.3	11.9	75.5	11.9	0.3	0	0
	Evidence/Elaboration	30,323	5,520	0.5	11.7	75.6	11.7	0.5	0	0
	Conventions	30,323	5,520	0.2	6.5	86.6	6.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.9 Writing Prompt Rater Agreement Report - Grade 11 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,190	4,190	0.5	17.4	64.3	17.4	0.5	0	0
	Evidence/Elaboration	4,190	4,190	0.8	18.4	61.6	18.4	0.8	0	0
	Conventions	4,190	4,190	0.2	12.2	75.2	12.2	0.2	0	0
2	Purpose/Organization	4,166	4,166	0.6	16.5	65.8	16.5	0.6	0	0
	Evidence/Elaboration	4,166	4,166	0.9	17.2	63.8	17.2	0.9	0	0
	Conventions	4,166	4,166	0.2	14.5	70.7	14.5	0.2	0	0
3	Purpose/Organization	4,176	4,176	0.9	15.7	66.8	15.8	0.9	0	0
	Evidence/Elaboration	4,176	4,176	1.4	17.6	62	17.6	1.4	0	0
	Conventions	4,176	4,176	0.4	12.5	74.2	12.5	0.4	0	0
4	Purpose/Organization	4,172	4,172	0.7	17.4	63.7	17.5	0.7	0	0
	Evidence/Elaboration	4,172	4,172	1	17.4	63.2	17.4	1	0	0
	Conventions	4,172	4,172	0.2	13.2	73.2	13.3	0.2	0	0
5	Purpose/Organization	4,180	4,180	0.6	16.6	65.5	16.6	0.6	0	0
	Evidence/Elaboration	4,180	4,180	1.1	19.1	59.5	19.1	1.1	0	0
	Conventions	4,180	4,180	0.2	12.6	74.4	12.6	0.2	0	0
6	Purpose/Organization	4,180	4,180	0.6	16.6	65.8	16.6	0.6	0	0
	Evidence/Elaboration	4,180	4,180	1.1	17.8	62.3	17.8	1.1	0	0
	Conventions	4,180	4,180	0.1	10.9	78	10.9	0.1	0	0
Paper Essay	Purpose/Organization	26,893	4,886	0.4	8.8	81.6	8.8	0.4	0	0
	Evidence/Elaboration	26,893	4,886	0.7	9.2	80.2	9.2	0.7	0	0
	Conventions	26,893	4,886	0.2	4.5	90.5	4.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix K.1—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 3 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post- equated	
1	1	835	-1.77	-1.36	-1.55	-0.19
2	2	837	-0.41	-0.14	-0.18	-0.04
3	3	8708	-2.08	-1.37	-1.85	-0.49
4	4	8710	-1.51	-1.13	-1.28	-0.15
5	5	8711	-0.33	-0.14	-0.11	0.04
6	6	9330	-0.49	-0.12	-0.26	-0.15
7	7	9331	0.40	0.54	0.63	0.09
8	8	9333	0.08	0.45	0.31	-0.14
9	9	9335	-0.87	-0.21	-0.65	-0.44
10	10	9336	-0.32	0.05	-0.10	-0.14
11	11	9347	0.05	0.16	0.28	0.12
12	12	9349	0.59	1.17	0.81	-0.35
13	13	9353	-0.12	0.10	0.10	0.01
14	14	9355	-0.59	-0.24	-0.36	-0.12
15	15	9356	1.08	0.96	1.30	0.34
16	16	9357	0.60	0.62	0.82	0.20
17	17	9359	0.05	0.20	0.28	0.08
18	18	9377	-0.04	0.12	0.19	0.07
19	19	9378	-1.31	-1.33	-1.09	0.24
19	20	9378	-0.58	-0.23	-0.36	-0.13
20	21	9379	-1.26	-0.68	-1.03	-0.35
21	22	9380	-1.11	-0.91	-0.88	0.03
22	23	9398	-0.28	0.22	-0.06	-0.27
23	24	9400	0.27	0.44	0.50	0.05
24	25	9401	0.87	1.23	1.09	-0.14
25	26	9402	1.43	1.71	1.66	-0.05
26	27	9404	-0.49	-0.34	-0.26	0.08
27	28	9414	-0.26	-0.42	-0.03	0.39
28	29	9418	0.48	0.51	0.71	0.20
29	30	9422	-1.10	-0.87	-0.88	-0.01
30	31	9687	1.16	1.38	1.38	0.00
31	32	9690	-1.44	-1.36	-1.21	0.15
32	33	9691	-0.60	-0.64	-0.38	0.27
33	34	9692	-0.83	-0.57	-0.60	-0.03
34	35	9694	1.48	1.36	1.71	0.35
35	36	9697	-0.49	-0.27	-0.27	0.00
36	37	9698	-1.34	-1.31	-1.12	0.19
37	38	9699	-0.53	-0.05	-0.30	-0.25
38	39	9700	1.01	0.87	1.24	0.36
39	40	10268	-0.03	0.44	0.20	-0.24
40	41	10632	0.25	0.54	0.48	-0.06
41	42	10634	0.95	0.72	1.18	0.46
Mean			-0.23	0.00		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.23	

Appendix K.2—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 4 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post-equated	
1	1	1122	0.41	1.09	0.41	-0.68
2	2	1124	-1.62	-0.21	-1.62	-1.41
3	3	9382	-2.14	-0.96	-2.14	-1.18
4	4	9386	-0.49	0.74	-0.49	-1.23
5	5	9387	0.52	1.21	0.52	-0.69
6	6	9388	-1.29	-0.24	-1.29	-1.05
7	7	9389	-0.92	0.23	-0.92	-1.15
8	8	9390	-1.85	-0.48	-1.85	-1.37
9	9	9391	-0.69	0.38	-0.69	-1.07
10	10	9392	-1.64	-0.36	-1.64	-1.28
11	11	9395	-1.24	-0.16	-1.24	-1.08
12	12	9397	-0.15	0.98	-0.15	-1.12
13	13	9425	-1.60	-0.60	-1.60	-1.00
14	14	9426	-0.03	1.05	-0.03	-1.08
15	15	9428	-0.16	0.68	-0.16	-0.84
16	16	9429	-3.17	-2.16	-3.17	-1.01
16	17	9429	-0.38	0.38	-0.38	-0.76
17	18	9431	-0.48	0.56	-0.48	-1.05
18	19	9437	-0.99	-0.19	-0.99	-0.80
19	20	9438	0.80	1.18	0.80	-0.38
20	21	9439	1.13	1.50	1.13	-0.37
21	22	9446	-1.73	-1.01	-1.73	-0.72
22	23	9450	-1.13	-0.24	-1.13	-0.89
23	24	9451	0.22	0.48	0.22	-0.26
24	25	9595	-1.38	-0.86	-1.38	-0.52
25	26	9596	-0.31	0.61	-0.31	-0.91
26	27	9597	-1.05	-0.16	-1.05	-0.89
27	28	9598	-0.97	-0.18	-0.97	-0.79
28	29	9603	1.77	2.60	1.77	-0.83
29	30	9616	-1.15	0.09	-1.15	-1.24
30	31	9899	-0.43	0.47	-0.43	-0.90
31	32	9900	1.77	2.27	1.77	-0.50
32	33	9902	-2.17	-1.22	-2.17	-0.95
33	34	9903	-0.40	0.42	-0.40	-0.81
34	35	9906	-1.50	-0.53	-1.50	-0.96
35	36	10263	-1.68	-0.61	-1.68	-1.07
36	37	10266	-1.70	-0.61	-1.70	-1.10
37	38	10277	0.16	0.85	0.16	-0.69
38	39	10278	1.17	1.78	1.17	-0.61
39	40	10644	-1.01	-0.09	-1.01	-0.92
40	41	10645	-3.05	-0.90	-3.05	-2.15
40	42	10645	-0.84	0.98	-0.84	-1.82
41	43	10647	-1.81	-0.82	-1.81	-1.00
Mean			-0.77	0.18		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.96	

Appendix K.3—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 5 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post- equated	
1	1	9290	-1.38	-0.74	-0.71	0.03
2	2	9291	-1.82	-1.07	-1.14	-0.07
2	3	9291	0.02	0.37	0.70	0.32
3	4	9292	-1.76	-1.15	-1.08	0.07
3	5	9292	-0.11	0.38	0.57	0.19
4	6	9294	-1.72	-1.16	-1.04	0.12
5	7	9298	-1.50	-0.40	-0.82	-0.42
6	8	9299	-0.81	-0.09	-0.14	-0.05
7	9	9302	-0.28	0.11	0.40	0.28
8	10	9303	-0.53	-0.18	0.14	0.33
9	11	9304	-0.15	0.27	0.53	0.26
10	12	9305	-1.74	-1.23	-1.07	0.16
11	13	9306	-1.01	-0.13	-0.34	-0.20
12	14	9307	-1.35	-0.45	-0.68	-0.23
13	15	9308	-1.00	0.28	-0.32	-0.61
14	16	9309	2.72	3.37	3.40	0.03
15	17	9311	1.43	2.27	2.11	-0.16
16	18	9312	2.05	3.51	2.72	-0.78
17	19	9320	0.50	1.44	1.18	-0.26
18	20	9599	-0.56	0.02	0.12	0.10
19	21	9600	-0.38	0.74	0.29	-0.45
20	22	9601	-0.88	-0.39	-0.20	0.19
21	23	9751	-0.91	-0.17	-0.24	-0.07
22	24	9752	0.52	1.41	1.20	-0.21
23	25	9753	0.14	0.90	0.82	-0.08
24	26	9754	2.56	3.10	3.24	0.13
25	27	9755	0.57	1.32	1.25	-0.07
26	28	9757	-0.76	0.05	-0.08	-0.14
27	29	9758	-1.16	-0.78	-0.48	0.30
28	30	9762	-0.62	-0.20	0.05	0.25
29	31	9763	-0.20	-0.06	0.48	0.54
30	32	9765	0.73	1.07	1.41	0.34
31	33	9767	-0.61	0.25	0.07	-0.18
32	34	9783	-1.54	-0.77	-0.86	-0.09
33	35	9784	0.06	0.44	0.74	0.30
34	36	9808	-0.48	0.24	0.20	-0.04
35	37	9809	-0.43	0.23	0.24	0.01
36	38	9833	0.13	0.70	0.81	0.11
37	39	9842	-0.05	0.45	0.63	0.18
38	40	10264	1.95	2.61	2.63	0.02
39	41	10659	-1.50	-0.97	-0.83	0.14
40	42	10661	-2.88	-2.09	-2.20	-0.11
40	43	10661	-1.34	-0.43	-0.67	-0.24
41	44	10662	-1.69	-1.09	-1.01	0.07
Mean			-0.40	0.27		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.68	

Appendix K.4—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 6 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post- equated	
1	1	9083	-2.86	-2.27	-2.37	-0.10
1	2	9083	0.21	0.49	0.69	0.20
2	3	9084	-1.01	-0.52	-0.53	-0.01
2	4	9084	1.46	1.80	1.95	0.15
3	5	9108	-1.80	-1.19	-1.32	-0.13
3	6	9108	0.76	1.12	1.24	0.12
4	7	9109	-1.98	-1.50	-1.49	0.01
4	8	9109	0.14	0.22	0.62	0.40
5	9	9130	-0.94	-0.93	-0.45	0.48
6	10	9131	0.39	0.72	0.88	0.16
7	11	9134	-0.73	-0.06	-0.24	-0.18
8	12	9135	0.04	0.45	0.53	0.08
9	13	9138	-1.96	-1.66	-1.48	0.18
10	14	9139	-1.39	-0.50	-0.91	-0.41
11	15	9140	-0.19	0.47	0.29	-0.17
12	16	9142	1.11	1.72	1.59	-0.13
13	17	9143	-0.10	0.53	0.39	-0.14
14	18	9145	-1.44	-0.50	-0.96	-0.46
15	19	9153	-0.90	-0.68	-0.42	0.27
16	20	9168	-0.13	0.14	0.35	0.21
17	21	9169	0.94	0.99	1.43	0.44
18	22	9262	0.70	1.28	1.19	-0.09
19	23	9263	-1.00	-0.14	-0.51	-0.37
20	24	9264	0.53	0.81	1.01	0.20
21	25	9266	-1.53	-0.90	-1.04	-0.14
22	26	9267	-1.08	-0.48	-0.59	-0.11
23	27	9268	1.60	2.10	2.08	-0.02
24	28	9272	-0.49	-0.16	-0.01	0.15
25	29	9797	0.49	1.00	0.98	-0.02
26	30	9798	-0.37	-0.44	0.11	0.55
27	31	9799	-0.65	-0.15	-0.17	-0.02
28	32	9800	0.73	1.44	1.21	-0.23
29	33	9802	-0.06	0.35	0.42	0.07
30	34	9803	-2.25	-1.75	-1.77	-0.02
31	35	9804	-0.98	-0.37	-0.50	-0.13
32	36	9807	0.68	1.33	1.17	-0.16
33	37	9865	0.17	1.65	0.65	-1.00
34	38	9866	-0.41	0.03	0.07	0.04
35	39	9867	0.40	0.59	0.88	0.30
36	40	9869	-1.77	-1.52	-1.28	0.24
37	41	9870	-0.67	-0.21	-0.19	0.02
38	42	9872	1.39	1.75	1.88	0.13
39	43	9889	-0.44	0.10	0.05	-0.06
40	44	10265	-1.53	-0.74	-1.05	-0.31
41	45	10280	-0.34	0.13	0.15	0.01
Mean			-0.38	0.10		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.48	

Appendix K.5—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 7 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post- equated	
1	1	9103	-0.25	0.13	0.33	0.20
2	2	9104	-1.80	-1.26	-1.22	0.04
2	3	9104	0.04	0.29	0.62	0.33
3	4	9105	-0.72	-0.33	-0.14	0.18
4	5	9128	0.70	1.44	1.29	-0.15
5	6	9146	-1.78	-1.18	-1.20	-0.01
6	7	9147	-1.13	-0.16	-0.54	-0.38
7	8	9152	-0.83	-0.20	-0.25	-0.05
8	9	9176	0.85	1.69	1.43	-0.26
9	10	9177	-0.04	0.73	0.55	-0.18
10	11	9189	0.55	0.79	1.13	0.35
11	12	9220	-0.81	-0.34	-0.23	0.12
12	13	9610	0.33	0.55	0.91	0.36
13	14	9611	-0.14	0.82	0.44	-0.38
14	15	9614	-1.60	-0.85	-1.02	-0.17
15	16	9709	-0.83	0.76	-0.24	-1.01
16	17	9711	-0.29	-0.12	0.30	0.42
17	18	9713	-0.35	0.47	0.23	-0.24
18	19	9750	-0.76	0.51	-0.18	-0.68
19	20	9777	0.19	0.64	0.77	0.13
20	21	9778	-0.16	0.34	0.43	0.09
21	22	9779	-0.26	0.16	0.32	0.16
22	23	9781	-0.36	0.09	0.23	0.14
23	24	9785	0.78	1.13	1.36	0.22
24	25	9786	0.45	1.07	1.03	-0.03
25	26	9787	0.00	0.41	0.59	0.18
26	27	9789	-0.83	-0.45	-0.25	0.20
27	28	9791	-0.32	-0.14	0.27	0.41
28	29	9793	1.04	0.73	1.63	0.90
29	30	9811	-0.96	-0.43	-0.38	0.05
30	31	9817	-0.69	0.01	-0.10	-0.12
31	32	9962	0.01	0.53	0.59	0.06
32	33	9982	-0.21	0.62	0.37	-0.25
33	34	10274	1.12	1.48	1.71	0.22
34	35	10613	0.59	1.53	1.18	-0.35
35	36	10619	-0.69	0.31	-0.11	-0.41
36	37	10621	0.45	1.01	1.04	0.03
37	38	10623	-0.28	0.48	0.30	-0.18
38	39	10656	-0.58	0.05	0.01	-0.04
39	40	10657	-1.89	-1.51	-1.31	0.21
39	41	10657	-0.70	-0.21	-0.12	0.09
40	42	10658	-1.34	-0.86	-0.76	0.11
40	43	10658	1.24	1.58	1.82	0.24
41	44	10695	-1.10	0.02	-0.52	-0.53
Mean			-0.30	0.28		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.58	

Appendix K.6—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 8 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post- equated	
1	1	8965	-1.72	-1.48	-1.59	-0.11
2	2	8967	-0.90	-0.36	-0.76	-0.40
3	3	8971	1.02	0.75	1.16	0.41
4	4	8973	-0.35	0.13	-0.22	-0.35
5	5	9014	-1.13	-0.91	-0.99	-0.08
6	6	9015	1.08	1.08	1.21	0.13
7	7	9017	-0.23	-0.44	-0.09	0.35
8	8	9018	0.68	0.56	0.81	0.25
9	9	9019	-0.09	-0.27	0.05	0.32
10	10	9020	0.96	1.45	1.09	-0.35
11	11	9022	0.86	1.22	0.99	-0.23
12	12	9023	0.82	0.69	0.96	0.26
13	13	9024	-0.67	-0.55	-0.54	0.01
14	14	9026	0.37	0.32	0.51	0.19
15	15	9028	0.04	-0.34	0.18	0.52
16	16	9044	0.76	0.94	0.89	-0.05
17	17	9046	-0.96	-1.12	-0.82	0.29
18	18	9076	-0.78	-0.28	-0.64	-0.37
19	19	9077	-1.93	-1.65	-1.80	-0.15
19	20	9077	0.20	0.28	0.33	0.05
20	21	9078	-1.07	-0.58	-0.94	-0.36
20	22	9078	0.67	0.99	0.81	-0.18
21	23	9113	-0.32	-0.30	-0.18	0.11
22	24	9115	0.61	0.62	0.75	0.13
23	25	9116	-0.42	-0.20	-0.28	-0.08
24	26	9170	0.37	0.65	0.50	-0.14
25	27	9171	1.71	1.74	1.84	0.10
26	28	9172	0.37	0.49	0.51	0.01
27	29	9230	-1.29	-1.16	-1.15	0.00
28	30	9240	-0.49	-0.25	-0.35	-0.10
29	31	9253	-0.76	-0.86	-0.62	0.24
30	32	9254	-0.46	-0.66	-0.32	0.34
31	33	9255	0.57	0.88	0.70	-0.18
32	34	9256	-0.38	0.08	-0.25	-0.33
33	35	9259	0.51	0.51	0.65	0.13
34	36	9728	-1.79	-1.75	-1.66	0.09
34	37	9728	-0.34	-0.25	-0.20	0.05
35	38	9729	-1.86	-1.61	-1.73	-0.12
35	39	9729	0.13	0.26	0.27	0.01
36	40	9880	-0.18	-0.09	-0.05	0.04
37	41	9881	0.43	0.50	0.56	0.06
38	42	9883	-1.62	-1.27	-1.48	-0.22
39	43	9885	-0.81	-0.46	-0.68	-0.22
40	44	10626	1.13	1.04	1.26	0.23
41	45	10639	2.78	3.21	2.92	-0.30
Mean			-0.10	0.04		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.13	

Appendix K.7—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 9 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post-equated	
1	1	1444	-0.65	-0.15	-0.42	-0.26
2	2	1446	-0.26	0.28	-0.03	-0.31
3	3	8948	-0.05	0.11	0.19	0.08
4	4	8951	-1.12	-0.68	-0.89	-0.21
4	5	8951	1.21	1.34	1.45	0.11
5	6	8987	-0.81	-0.88	-0.57	0.30
6	7	8989	-1.24	-0.80	-1.01	-0.21
7	8	8990	0.28	0.11	0.51	0.40
8	9	9003	0.07	-0.09	0.30	0.40
9	10	9004	-0.23	-0.12	0.01	0.13
10	11	9005	0.20	0.59	0.43	-0.16
11	12	9006	-0.16	0.22	0.07	-0.14
12	13	9007	-0.38	-0.08	-0.14	-0.07
13	14	9012	0.16	0.42	0.40	-0.02
14	15	9032	0.58	0.67	0.82	0.15
15	16	9033	-0.39	-0.26	-0.15	0.10
16	17	9034	0.95	0.96	1.18	0.22
17	18	9038	0.32	0.49	0.55	0.07
18	19	9040	0.77	0.40	1.00	0.60
19	20	9041	1.61	1.65	1.84	0.19
20	21	9042	0.36	0.34	0.59	0.25
21	22	9043	0.14	0.26	0.37	0.12
22	23	9047	0.81	1.43	1.04	-0.39
23	24	9048	-0.62	-0.32	-0.39	-0.07
24	25	9050	1.15	1.01	1.38	0.37
25	26	9051	-0.03	0.22	0.20	-0.01
26	27	9053	-1.07	-0.65	-0.84	-0.19
27	28	9058	0.13	0.72	0.36	-0.37
28	29	9060	1.04	1.58	1.28	-0.31
29	30	9063	-0.65	-0.28	-0.42	-0.14
30	31	9064	0.98	0.80	1.21	0.41
31	32	9065	0.06	0.46	0.29	-0.17
32	33	9066	0.14	0.40	0.37	-0.03
33	34	9069	0.79	0.83	1.02	0.19
34	35	9734	-0.62	-0.66	-0.39	0.28
35	36	9735	-1.25	-1.05	-1.01	0.03
35	37	9735	1.35	1.26	1.58	0.32
36	38	9736	-0.52	-0.44	-0.29	0.15
37	39	9737	-1.64	-1.52	-1.41	0.11
38	40	10595	-1.24	-0.57	-1.01	-0.44
39	41	10596	-1.86	-1.03	-1.63	-0.60
40	42	10597	-1.54	-0.56	-1.31	-0.74
41	43	10603	0.14	0.53	0.37	-0.16
42	44	11097	-0.68	-0.29	-0.45	-0.16
43	45	11098	0.18	0.22	0.42	0.20
Mean			-0.08	0.15		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.23	

Appendix K.8—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 10 ELA

No. Items	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post-equated	
1	1	8757	-1.50	-1.10	-1.37	-0.27
2	2	8758	0.20	0.36	0.33	-0.03
3	3	8759	-1.30	-1.33	-1.18	0.16
3	4	8759	-0.35	-0.25	-0.22	0.04
4	5	8760	0.74	1.03	0.86	-0.16
5	6	8761	-0.23	0.13	-0.10	-0.23
6	7	8762	0.45	0.51	0.58	0.07
7	8	8764	-0.80	-0.43	-0.67	-0.24
7	9	8764	0.99	1.07	1.12	0.05
8	10	8785	-1.37	-0.99	-1.24	-0.25
9	11	8788	-0.92	-0.60	-0.80	-0.20
10	12	8795	1.40	1.59	1.53	-0.06
11	13	8815	0.43	0.68	0.56	-0.12
12	14	8816	0.52	0.51	0.65	0.14
13	15	8818	0.28	0.82	0.41	-0.41
14	16	8819	0.50	0.97	0.62	-0.34
15	17	8821	0.06	-0.24	0.19	0.42
16	18	8844	2.96	2.67	3.08	0.42
17	19	8925	-0.49	0.10	-0.36	-0.46
18	20	8926	0.30	0.64	0.42	-0.21
19	21	8972	0.33	0.41	0.46	0.05
20	22	9623	-0.46	-0.24	-0.33	-0.09
21	23	9624	0.84	1.15	0.97	-0.18
22	24	9626	0.27	0.40	0.40	0.00
23	25	9627	-0.11	-0.10	0.02	0.11
24	26	9630	0.41	0.47	0.54	0.07
25	27	9769	-0.11	-0.24	0.02	0.26
26	28	9771	-0.57	-0.32	-0.44	-0.12
27	29	9772	-1.40	-1.81	-1.27	0.54
28	30	9773	0.73	0.69	0.86	0.16
29	31	9774	-0.21	-0.07	-0.08	-0.01
30	32	9821	-0.09	-0.07	0.03	0.10
31	33	9822	-0.68	-1.19	-0.55	0.65
32	34	9824	0.70	1.01	0.83	-0.18
33	35	9825	-0.65	-0.37	-0.53	-0.16
34	36	9826	-0.57	-0.34	-0.44	-0.10
35	37	9827	0.21	0.42	0.33	-0.08
36	38	9829	-0.15	-0.30	-0.02	0.28
37	39	9830	1.69	1.91	1.82	-0.09
38	40	9831	-1.26	-1.27	-1.14	0.13
39	41	9837	-0.51	-0.32	-0.39	-0.07
40	42	9839	-0.20	-0.06	-0.07	0.00
41	43	9840	0.37	0.25	0.49	0.24
42	44	9887	1.86	1.98	1.99	0.01
43	45	9888	-0.16	-0.20	-0.03	0.17
Mean			0.05	0.18		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				0.13	

Appendix K.9—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 11 ELA

No. Items	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post-equated	
1	1	421	-0.11	-0.04	-0.19	-0.15
2	2	422	1.15	1.42	1.07	-0.35
3	3	8753	-1.22	-1.40	-1.30	0.10
4	4	8754	-1.20	-1.30	-1.28	0.03
5	5	8755	-0.41	-0.31	-0.49	-0.18
5	6	8755	1.74	1.75	1.66	-0.09
6	7	8769	0.47	0.39	0.39	0.01
7	8	8770	-1.49	-1.41	-1.57	-0.16
7	9	8770	0.33	0.32	0.25	-0.07
8	10	8771	-1.21	-1.37	-1.28	0.09
9	11	8781	0.42	0.76	0.34	-0.41
10	12	8783	-0.60	-1.12	-0.68	0.44
11	13	8784	-0.18	-0.69	-0.26	0.44
12	14	8791	1.39	0.81	1.31	0.50
13	15	8792	-0.14	-0.33	-0.21	0.11
14	16	8794	0.68	0.78	0.60	-0.17
15	17	8796	0.14	-0.04	0.06	0.10
16	18	8797	-0.28	-0.29	-0.36	-0.07
17	19	8798	0.25	0.18	0.17	-0.01
18	20	8799	-0.44	-0.48	-0.52	-0.04
19	21	8800	-0.70	-0.86	-0.77	0.09
20	22	8805	0.36	-0.09	0.28	0.37
21	23	8807	0.12	0.06	0.04	-0.03
22	24	8808	-0.35	-0.60	-0.43	0.17
23	25	8809	0.19	0.01	0.12	0.10
24	26	8834	-0.02	-0.47	-0.10	0.38
25	27	8837	1.15	1.16	1.08	-0.08
26	28	8841	0.14	-0.13	0.06	0.19
27	29	8843	0.90	0.91	0.82	-0.09
28	30	8846	0.30	-0.06	0.22	0.28
29	31	8856	0.62	0.22	0.54	0.32
30	32	8861	0.34	0.32	0.26	-0.06
31	33	8862	0.84	0.76	0.77	0.00
32	34	8865	0.65	0.44	0.57	0.13
33	35	8867	0.26	0.13	0.18	0.05
34	36	8871	0.12	0.02	0.04	0.02
35	37	8879	0.60	1.10	0.53	-0.58
36	38	8880	0.05	-0.24	-0.03	0.21
37	39	8881	-0.29	-0.69	-0.37	0.32
38	40	8884	0.34	0.31	0.26	-0.04
39	41	9852	-0.77	-0.25	-0.84	-0.59
40	42	9853	-0.67	-0.18	-0.75	-0.57
41	43	9856	-0.05	-0.02	-0.13	-0.11
42	44	9858	1.07	1.04	0.99	-0.05
43	45	9860	-0.26	0.21	-0.34	-0.55
Mean			0.09	0.02		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				-0.08	

Appendix K.10 — Field Test Analysis: Linking Operational Parameters to the Bank Scale—Grade 3 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10384	-2.07	-1.18	-1.36	-0.18
10391	-1.21	-0.23	-0.50	-0.26
10392	-0.08	0.63	0.63	-0.01
10395	-2.75	-1.62	-2.04	-0.42
10396	-0.33	0.10	0.38	0.28
10398	-0.26	0.81	0.45	-0.36
10399	0.18	1.28	0.89	-0.40
10400	-0.97	-0.63	-0.26	0.37
10402	-2.74	-1.93	-2.03	-0.10
10403	-1.29	-0.64	-0.58	0.06
10409	-2.58	-1.75	-1.87	-0.12
10411	-2.33	-1.43	-1.62	-0.19
10416	-0.39	0.72	0.32	-0.40
10418	-0.93	-0.36	-0.22	0.14
10425	-2.25	-1.52	-1.54	-0.01
10427	0.36	0.91	1.07	0.16
10430	1.61	2.22	2.32	0.11
10431	-1.79	-1.08	-1.08	0.00
10433	0.63	1.21	1.34	0.13
10434	-1.46	-0.86	-0.75	0.11
10436	-1.21	-0.16	-0.50	-0.35
10438	-3.05	-1.82	-2.34	-0.53
10439	-1.72	-0.67	-1.01	-0.35
10443	-3.04	-2.22	-2.33	-0.11
10446	0.58	1.07	1.29	0.22
10448	0.39	1.01	1.10	0.08
10450	-0.37	-0.06	0.34	0.40
10453	-4.10	-2.99	-3.39	-0.40
10454	1.57	2.12	2.28	0.17
10455	1.14	1.71	1.85	0.15
10466	0.12	0.78	0.83	0.05
10470	1.53	2.08	2.24	0.16
10671	-0.32	0.27	0.39	0.12
10683	0.72	1.28	1.43	0.15
10685	0.99	1.50	1.70	0.20
10687	0.19	0.79	0.90	0.10
11120	0.42	1.12	1.13	0.01
8461	-1.31	-0.87	-0.60	0.27
8463	-1.30	-0.57	-0.59	-0.02
8481	0.11	0.72	0.82	0.10
8483	-0.73	-0.85	-0.02	0.83
9455	-0.78	-0.11	-0.07	0.04
9460	-2.39	-1.49	-1.68	-0.18
9464	0.61	1.20	1.32	0.12
9469	0.86	1.71	1.57	-0.14
mean	-0.71	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.71	

Appendix K.11 — Field Test Analysis: Linking Operational Parameters to the Bank Scale—Grade 4 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10705	-0.48	0.07	0.15	0.09
10708	-1.00	-0.73	-0.36	0.37
10710	-1.07	-0.18	-0.44	-0.26
10715	0.11	0.43	0.74	0.31
10716	0.47	1.13	1.10	-0.03
10718	-0.60	-0.12	0.03	0.15
10719	-0.67	0.01	-0.04	-0.04
10724	0.50	1.02	1.13	0.11
10727	1.52	2.13	2.16	0.03
10728	0.66	1.07	1.29	0.22
10729	-1.61	-0.93	-0.97	-0.04
10730	-1.56	-0.50	-0.93	-0.43
10731	-1.41	-1.09	-0.78	0.31
10739	0.05	0.47	0.68	0.21
10744	-1.13	0.11	-0.50	-0.61
10750	-1.22	-0.28	-0.59	-0.31
10751	-0.97	-0.46	-0.34	0.12
10752	-0.19	0.41	0.44	0.03
10753	0.23	1.15	0.86	-0.29
10754	0.80	1.36	1.43	0.07
10763	-0.84	-0.27	-0.20	0.07
10766	-2.10	-1.15	-1.47	-0.32
10771	-1.61	-0.77	-0.98	-0.21
10772	-3.24	-2.20	-2.60	-0.40
10779	-0.29	0.21	0.34	0.14
10780	1.62	2.22	2.25	0.03
10780	1.87	2.19	2.50	0.31
10781	0.05	0.44	0.68	0.24
10782	0.83	1.67	1.46	-0.21
10783	-1.21	-0.74	-0.58	0.16
10826	-2.13	0.59	-1.49	-2.09
10826	0.17	-1.20	0.81	2.01
10827	-1.50	-0.66	-0.86	-0.21
8493	-1.43	-1.00	-0.80	0.20
8497	-2.02	-1.07	-1.38	-0.31
8499	-2.75	-1.91	-2.11	-0.21
8501	-0.13	0.11	0.50	0.39
8505	-2.41	-1.65	-1.77	-0.12
9452	-0.29	0.49	0.35	-0.15
9456	-3.71	-2.94	-3.08	-0.14
9465	0.00	0.63	0.64	0.00
9467	0.29	0.79	0.92	0.14
9470	1.33	1.88	1.96	0.08
9471	0.69	1.10	1.33	0.23
9474	-0.81	0.01	-0.18	-0.19
9475	1.56	2.10	2.19	0.09
9502	-0.31	-0.11	0.33	0.44
mean	-0.55	0.08		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.63	

Appendix K.12— Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 5 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10788	-0.26	0.37	0.28	-0.09
10789	-3.37	-2.75	-2.84	-0.09
10790	0.25	1.20	0.79	-0.42
10791	0.36	1.18	0.90	-0.28
10795	-1.93	-1.35	-1.40	-0.04
10796	-0.25	0.01	0.28	0.27
10798	-0.72	-0.34	-0.18	0.16
10799	-1.13	-0.43	-0.59	-0.16
10800	0.58	1.29	1.12	-0.18
10803	-0.75	-0.05	-0.22	-0.17
10805	-1.02	-0.26	-0.49	-0.23
10808	0.42	1.00	0.96	-0.04
10811	-1.65	-1.07	-1.11	-0.05
10813	0.88	1.58	1.42	-0.16
10816	-0.51	0.67	0.02	-0.65
10817	2.64	1.98	3.18	1.19
10817	1.33	3.21	1.87	-1.34
10820	1.09	1.16	1.62	0.46
10824	-0.71	-0.31	-0.18	0.13
10829	-2.13	-1.20	-1.59	-0.40
10832	-1.14	-0.66	-0.61	0.05
10833	-1.53	-0.95	-1.00	-0.05
10835	1.59	1.82	2.12	0.30
10836	-1.29	-0.63	-0.75	-0.12
10839	0.10	0.66	0.63	-0.03
10840	0.43	1.11	0.96	-0.15
10848	-1.35	-1.03	-0.81	0.21
10849	0.62	1.10	1.15	0.05
10850	-0.34	0.38	0.20	-0.18
10851	0.39	1.02	0.92	-0.09
10863	0.09	0.38	0.63	0.24
10868	3.10	3.74	3.64	-0.10
10869	-0.32	0.24	0.21	-0.02
10872	-1.30	-0.89	-0.77	0.12
10872	-1.02	-1.00	-0.48	0.52
10875	-1.23	-1.09	-0.69	0.40
11107	-1.00	-0.43	-0.47	-0.04
8521	-1.71	-1.11	-1.18	-0.07
8525	-1.12	-0.93	-0.59	0.35
8527	-1.22	-0.87	-0.69	0.18
8535	-1.51	-1.22	-0.97	0.25
8539	-1.71	-1.34	-1.17	0.17
9476	0.82	1.56	1.36	-0.21
9485	1.23	1.66	1.76	0.10
9486	-0.39	0.35	0.15	-0.20
9487	0.58	0.66	1.11	0.46
9716	-2.84	-2.23	-2.30	-0.07
mean	-0.40	0.13		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.54	

Appendix K.13 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 6 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10048	-0.08	0.58	0.41	-0.18
10049	-0.58	0.22	-0.09	-0.32
10051	-0.85	-0.05	-0.36	-0.31
10052	-0.34	-0.69	0.15	0.84
10053	-1.27	-0.52	-0.78	-0.26
10057	-1.69	-0.23	-1.20	-0.97
10060	0.72	1.14	1.21	0.06
10064	-0.61	-0.12	-0.12	0.00
10067	-4.21	-3.16	-3.73	-0.56
10070	-0.19	1.07	0.30	-0.78
10071	0.08	0.50	0.56	0.06
10076	0.34	0.80	0.83	0.03
10078	0.56	0.85	1.05	0.21
10079	0.33	0.56	0.82	0.26
10083	0.43	0.82	0.92	0.09
10088	0.15	0.47	0.63	0.17
10093	-0.45	-0.08	0.04	0.11
10094	0.55	0.90	1.04	0.14
10095	0.48	0.49	0.97	0.48
10096	0.44	1.12	0.93	-0.19
10103	-1.42	-1.03	-0.93	0.10
10106	-2.36	-1.48	-1.87	-0.39
10107	-2.32	-1.73	-1.83	-0.10
10108	-1.16	-0.70	-0.67	0.04
10111	1.11	1.34	1.60	0.26
10113	-1.93	-1.75	-1.44	0.31
10115	-0.71	-0.83	-0.22	0.61
10117	0.40	0.87	0.89	0.03
10120	-0.02	0.41	0.47	0.06
10129	-2.79	-2.24	-2.30	-0.06
10137	-2.85	-2.11	-2.36	-0.25
10139	1.75	2.28	2.24	-0.04
10143	0.38	0.90	0.87	-0.03
10148	-0.90	-0.61	-0.42	0.19
10150	-0.53	-0.06	-0.04	0.01
10151	1.49	1.57	1.98	0.42
8549	-3.02	-2.29	-2.53	-0.24
8555	-0.67	-0.28	-0.18	0.11
8567	-1.14	-0.32	-0.66	-0.34
9491	-2.19	-1.65	-1.70	-0.05
9492	-1.91	-1.29	-1.42	-0.13
9496	-0.26	0.36	0.23	-0.14
9498	2.24	2.54	2.73	0.19
9512	0.89	1.25	1.38	0.12
9513	0.81	1.17	1.30	0.12
9718	1.85	1.98	2.34	0.36
9719	-1.78	-1.23	-1.29	-0.06
mean	-0.49	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.49	

Appendix K.14— Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 7 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10288	-1.96	-0.77	-0.97	-0.20
10290	-1.67	-0.49	-0.69	-0.20
10294	-0.87	0.36	0.11	-0.24
10298	-0.86	0.47	0.13	-0.34
10299	-1.48	-0.69	-0.49	0.20
10301	-1.53	-0.34	-0.54	-0.20
10302	0.18	1.28	1.16	-0.11
10303	-2.11	-0.78	-1.12	-0.34
10309	-0.01	0.71	0.98	0.27
10313	-1.81	-0.82	-0.83	-0.01
10317	0.83	2.01	1.81	-0.20
10318	-1.84	-0.63	-0.85	-0.22
10322	-1.23	-0.29	-0.24	0.05
10331	-0.60	0.15	0.38	0.24
10340	1.00	2.10	1.98	-0.12
10341	1.19	1.62	2.17	0.55
10344	0.07	1.18	1.06	-0.12
10347	0.83	1.92	1.82	-0.11
10349	-1.05	-0.12	-0.06	0.06
10351	0.35	1.25	1.34	0.09
10352	0.09	1.62	1.08	-0.54
10362	-1.52	-0.57	-0.54	0.03
10366	-0.61	0.41	0.38	-0.03
10366	-0.66	0.18	0.33	0.15
10369	-0.73	0.08	0.26	0.17
10371	-2.15	-1.55	-1.16	0.39
10374	-0.46	0.54	0.53	-0.01
10375	0.73	1.68	1.72	0.04
10378	-3.46	-2.14	-2.48	-0.33
10379	-0.52	0.17	0.47	0.30
10701	0.04	1.07	1.03	-0.04
8593	-1.19	-0.25	-0.20	0.05
8597	-0.91	0.06	0.08	0.02
8603	-1.43	-0.88	-0.44	0.44
8609	-0.62	0.32	0.36	0.05
8613	-2.04	-1.08	-1.05	0.03
8698	-3.04	-1.85	-2.05	-0.20
9508	0.89	1.81	1.88	0.07
9514	-0.09	0.81	0.90	0.09
9516	-0.70	0.01	0.29	0.28
9520	-0.72	0.27	0.27	0.00
mean	-0.77	0.21		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.99	

Appendix K.15 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 8 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10483	-1.15	-0.97	-0.67	0.30
10487	-0.60	-0.05	-0.13	-0.08
10488	-2.67	-2.00	-2.19	-0.19
10494	-0.56	0.10	-0.08	-0.19
10496	-0.82	-0.55	-0.34	0.21
10498	-1.39	-0.83	-0.92	-0.09
10507	-2.30	-1.61	-1.83	-0.22
10510	0.23	0.71	0.71	0.00
10513	-0.69	-0.72	-0.22	0.50
10518	0.97	1.56	1.45	-0.11
10520	-1.43	-1.02	-0.95	0.06
10523	2.21	2.94	2.69	-0.25
10525	-2.12	-1.58	-1.64	-0.06
10527	-0.36	-0.19	0.12	0.31
10528	-2.93	-2.49	-2.45	0.03
10530	-1.31	-0.77	-0.84	-0.06
10532	-0.31	0.17	0.17	-0.01
10534	-0.43	-0.28	0.05	0.32
10538	0.94	1.74	1.41	-0.33
10541	0.28	0.88	0.76	-0.12
10542	0.69	1.22	1.16	-0.06
10548	-0.12	-0.08	0.35	0.43
10554	-1.50	-0.78	-1.03	-0.24
10557	-3.64	-2.55	-3.17	-0.61
10561	-0.70	-0.24	-0.22	0.02
10562	-1.28	-1.01	-0.80	0.21
10564	-0.07	0.40	0.40	0.00
10567	-1.38	-0.72	-0.90	-0.18
10570	0.63	1.38	1.10	-0.28
10574	-0.74	-0.33	-0.27	0.07
10579	2.55	3.28	3.02	-0.26
10580	0.49	0.96	0.97	0.01
10581	-2.40	-1.86	-1.93	-0.07
10585	1.13	1.83	1.60	-0.22
10587	-0.04	0.42	0.44	0.01
10588	-1.54	-1.19	-1.06	0.12
8623	-2.11	-1.81	-1.64	0.17
8631	-1.70	-1.31	-1.23	0.08
8635	-1.31	-1.18	-0.83	0.35
8651	-0.60	0.03	-0.12	-0.16
9518	1.70	2.22	2.18	-0.04
9519	-0.52	-0.37	-0.04	0.33
9521	0.64	1.05	1.11	0.06
9522	2.03	2.62	2.51	-0.11
9525	-0.22	0.09	0.26	0.17
9525	-0.15	0.29	0.32	0.03
9527	-0.52	-0.23	-0.04	0.19
9532	0.26	0.78	0.73	-0.05
mean	-0.52	-0.04		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.48	

Appendix K.16 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Algebra I


Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10877	-1.76	-0.44	-0.40	0.04
10880	-0.04	1.22	1.32	0.09
10882	-0.55	0.86	0.81	-0.05
10887	-2.43	-1.05	-1.07	-0.02
10888	0.54	2.08	1.90	-0.18
10889	-1.53	-0.08	-0.17	-0.09
10895	-1.41	-0.49	-0.05	0.43
10896	-0.75	0.59	0.61	0.02
10897	-1.04	0.23	0.32	0.09
10905	-2.34	-1.19	-0.98	0.21
10906	-1.17	0.26	0.19	-0.07
10907	-2.48	-0.84	-1.12	-0.28
10934	-2.31	-1.27	-0.95	0.32
10935	-1.48	-0.12	-0.12	0.00
10941	0.16	1.51	1.52	0.01
10942	-1.26	0.11	0.10	-0.01
10943	-0.89	0.33	0.47	0.13
10945	-2.01	-0.70	-0.65	0.05
10951	-0.65	0.68	0.71	0.03
10953	-2.12	-0.94	-0.76	0.18
10960	-0.75	0.72	0.61	-0.11
10963	-2.24	-0.75	-0.88	-0.13
10966	-1.21	0.30	0.15	-0.15
10972	-1.46	-0.05	-0.10	-0.05
10973	-2.72	-1.42	-1.36	0.06
10974	-1.95	-0.27	-0.59	-0.32
10977	-1.78	-0.77	-0.42	0.35
10978	0.47	1.48	1.83	0.35
10988	-0.95	0.79	0.41	-0.37
10990	-0.70	0.81	0.66	-0.16
10993	-2.44	-1.24	-1.08	0.15
11004	-1.10	0.20	0.26	0.06
11011	-1.04	0.46	0.32	-0.14
11013	-1.26	0.21	0.10	-0.11
11044	-0.23	1.33	1.13	-0.20
11045	-1.45	-0.12	-0.09	0.03
11052	-2.62	-1.12	-1.26	-0.14
11058	-1.09	0.36	0.27	-0.08
9530	-0.67	0.43	0.69	0.26
9531	-0.39	1.14	0.97	-0.16
9533	-0.59	0.84	0.77	-0.07
9535	-1.40	-0.17	-0.04	0.13
9536	0.00	1.01	1.36	0.36
9543	-2.07	-0.67	-0.71	-0.04
9543	-1.75	-0.77	-0.39	0.37
9546	-0.56	0.82	0.80	-0.02
9705	-2.33	-0.48	-0.97	-0.49
9707	-3.53	-1.90	-2.17	-0.27
mean	-1.32	0.04		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		1.36	

Appendix K.17 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Geometry

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10910	-1.51	-1.16	-0.78	0.37
10912	-0.02	0.48	0.70	0.22
10919	-1.92	-1.12	-1.20	-0.08
10921	-2.55	-1.64	-1.82	-0.19
10924	-0.93	-0.19	-0.20	-0.01
10926	-0.25	0.71	0.47	-0.24
10930	0.01	0.46	0.73	0.27
10986	-1.88	-1.30	-1.15	0.14
10998	2.96	3.51	3.69	0.18
11007	-1.48	-0.32	-0.75	-0.43
11008	-0.19	0.26	0.54	0.28
11015	-2.38	-1.44	-1.65	-0.21
11016	-1.54	-0.67	-0.81	-0.14
11017	0.00	0.73	0.73	0.00
11018	-1.34	-0.87	-0.61	0.25
11026	-1.49	-0.62	-0.76	-0.15
11029	-0.86	-0.14	-0.13	0.01
11032	0.14	1.04	0.87	-0.17
11033	-1.79	-1.02	-1.07	-0.05
11034	-0.77	-0.08	-0.04	0.04
11035	-1.31	-0.67	-0.59	0.09
11036	-1.04	-0.42	-0.32	0.11
11037	-3.90	-3.37	-3.17	0.20
11040	-1.46	-0.44	-0.73	-0.29
11060	-0.25	0.63	0.47	-0.16
11061	-1.96	-1.39	-1.23	0.16
11065	0.29	0.94	1.01	0.07
11068	-1.37	-0.79	-0.65	0.14
11072	-1.30	-0.58	-0.57	0.01
11074	-0.15	0.64	0.58	-0.06
11078	-1.29	-0.26	-0.57	-0.30
11081	0.97	1.49	1.70	0.21
11085	-0.16	0.62	0.56	-0.06
11086	-1.79	-1.12	-1.07	0.05
11089	-1.50	-0.42	-0.77	-0.35
11092	0.21	0.87	0.93	0.06
11114	-0.85	-0.30	-0.12	0.18
9547	-1.99	-0.82	-1.26	-0.44
9551	1.34	2.21	2.06	-0.15
9554	0.22	0.75	0.94	0.19
9556	-0.95	-0.19	-0.22	-0.03
9560	-0.40	-0.27	0.32	0.59
9564	-0.29	0.48	0.44	-0.04
9575	2.11	2.96	2.84	-0.12
9581	-0.52	0.09	0.20	0.11
9592	0.78	1.56	1.51	-0.06
9722	0.29	1.20	1.01	-0.19
mean	-0.72	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.73	


Appendix K.18 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Algebra II

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10160	-1.45	-0.96	-0.66	0.30
10164	-0.11	0.70	0.69	-0.01
10168	0.89	1.54	1.69	0.15
10175	-2.09	-1.11	-1.29	-0.17
10176	0.02	1.21	0.82	-0.39
10177	-1.37	-0.67	-0.57	0.10
10180	1.33	2.15	2.12	-0.03
10182	0.58	1.38	1.38	-0.01
10187	-0.98	-0.52	-0.18	0.34
10192	-2.01	-1.25	-1.21	0.04
10199	0.15	1.20	0.95	-0.25
10200	-2.26	-1.20	-1.46	-0.26
10203	-1.82	-0.85	-1.02	-0.18
10204	-2.17	-1.42	-1.37	0.05
10206	-3.26	-2.25	-2.46	-0.21
10209	-1.46	-0.75	-0.66	0.09
10210	-2.59	-1.48	-1.80	-0.31
10214	-2.45	-1.96	-1.65	0.30
10215	-3.27	-2.31	-2.47	-0.16
10217	-2.10	-1.49	-1.30	0.19
10220	-0.25	0.55	0.54	-0.01
10223	-0.31	0.79	0.48	-0.30
10228	-0.24	0.50	0.55	0.05
10230	-1.16	-0.40	-0.36	0.04
10233	-1.61	-0.88	-0.81	0.07
10236	-1.69	-0.67	-0.90	-0.22
10237	-2.51	-1.98	-1.71	0.27
10240	-0.79	0.24	0.01	-0.23
10243	-1.26	-0.81	-0.47	0.35
10245	-0.14	0.75	0.66	-0.09
10249	0.27	1.26	1.07	-0.19
10255	0.38	1.53	1.18	-0.35
10256	0.25	0.88	1.05	0.17
10259	-0.89	-0.21	-0.09	0.12
10261	-0.29	0.33	0.51	0.18
11121	-1.84	-0.70	-1.04	-0.34
9548	1.49	1.89	2.28	0.39
9549	-0.76	-0.32	0.03	0.35
9567	0.19	0.83	0.98	0.15
9568	0.06	0.53	0.85	0.33
9570	-0.90	0.13	-0.10	-0.24
9573	-0.48	0.26	0.32	0.05
9577	0.43	1.57	1.23	-0.34
9578	-0.54	0.45	0.26	-0.19
9580	-0.85	-0.46	-0.06	0.41
9589	1.42	2.33	2.22	-0.11
9591	1.19	1.90	1.99	0.09
mean	-0.79	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.80	




AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Statistical Review Training for ADE




AIR
American Institutes for Research



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics


Statistical Review of

- Item Quality and Performance
 - Does the item behave the way it's supposed to behave?
- Item Difficulty
 - How hard is the item?
- Differential Item Functioning
 - Does the item behave




AIR
American Institutes for Research

2


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Quality

- Do highly skilled students perform better on the item than less skilled students?
- Correlation with Test – link between selecting a response option and doing well on the rest of the test
 - For key, + is good, – is bad
 - For distractors, – is good, + is bad


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

3


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Quality Flag Criteria

- Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items; (AB)
- Adjusted biserial correlations for multiple-choice item distractors is greater than .05; (ABD)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

4


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Difficulty

- How hard is the item?
- What percent of students answer item correctly?
- MC items – % of students selecting each response option
- Non-MC items – % of students achieving each score point


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

5


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Difficulty Flag Criteria

- Proportion correct value is less than .25 or greater than .95 for multiple-choice items, or greater than .95 for any single score point of a constructed-response item;
- Also known as p-value (P or CR_Prop)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

6


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Non-Modal Key

- A distractor is chosen by students more often than the key is chosen


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

7


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Non-Modal Key Flag Criteria

- The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items; (NMK)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

8


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Omit Rate

- Students do not provide a response


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

9


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Omit Rate Flag Criteria

- Omit rate is greater than .15;

 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

10


AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics


Differential Item Functioning

* Fair Items behave similarly across groups


* Probability of answering correctly is the same for all students of similar ability regardless of group membership

Subgroup Comparisons:

- Female/Male
- Non-Hispanic / Hispanic, Latino or Spanish origin
- Black, African American / White
- American Indian or Alaskan Native / White
- Asian / White
- Native Hawaiian or Other Pacific Islander / White
- Multiple ethnicities selected / White



AIR
AMERICAN INSTITUTES FOR RESEARCH

11



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Differential Item Functioning (DIF)

- Direction of possible bias
 - “–” item favors reference groups
 - “+” item favors focal group
- Severity of possible bias
 - “A” No statistical evidence of DIF
 - “B” Evidence for potential mild DIF
 - “C” Evidence for potential severe DIF
- “C” indicates that the item is more difficult for one group and should be reviewed carefully for bias



AIR
AMERICAN INSTITUTES FOR RESEARCH

12


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

DIF Flag Criteria

- Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF.
- Items are categorized as **positive DIF** (i.e., +A, +B, or +C), signifying that the item **favors the focal group** (e.g., African American/Black, Hispanic, or female), or
- **negative DIF** (i.e., -A, -B, or -C), signifying that the item **favors the reference group** (e.g., white or male).
- Items are flagged if their DIF statistics fall into the “C” category for any group, which indicates that the item shows **significant DIF** and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

13


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Content Expert Judgments

- Statistical information is important, but not a substitute for expert judges
- Items central to a learning standard may be difficult because a concept is not currently included in curriculum
- Items may show DIF because some concepts may be less likely to be covered in all area schools


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

14



Logistics

- Items can be found at the **Content and Fairness Data Review and Resolution** review level in the Arizona Assessment project in ITS
- The MDSs will be posted here on the sftp:
/files/AzMERIT/To ADE/Content Data Review/
- Please “PEND” any data comments in ITS

15

